UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Reversibility of Prion Misfolding by Constant-pH Molecular Dynamics Simulations

Diogo Ruivo dos Santos Vila Viçosa

Mestrado em Bioquímica
2010

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Reversibility of Prion Misfolding by Constant-pH Molecular Dynamics Simulations

Diogo Ruivo dos Santos Vila Viçosa
Tese orientada por Doutor António Baptista e Doutor Miguel Machuqueiro

Mestrado em Bioquímica
2010

Abstract

The prion protein (PrP) is the cause of a group of diseases known as Transmissible Spongiform Encephalopathies (TSEs). Creutzfeldt-Jakob and Bovine Spongiform Encephalopathy are examples of TSEs. The normal form of PrP (PrP^{C}) is monomeric and soluble, however, it can misfold into a pathogenic form (PrP^{Sc}). This last form has a high content of β-structures and can aggregate forming amyloid fibrils. The mechanism of conversion between PrP^{C} and PrP^{Sc} is not completely elucidated but it can be catalyzed by a PrP^{Sc} sample (protein-only hypothesis) or it can be induced by an external factor. The pH seems to be a factor that can induce the misfolding transition and it may occur in the endocytic pathway. The pH effect in the structure of PrP was studied recently in Molecular Simulation Group at ITQB [**1**] and an evident misfolding transition was observed in one simulation at pH 2. The main goal of the present work was to study the effects of a change in pH to 7 in several transient conformations of this simulation. To address this problem, we performed a total of 47 simulations, using our own Constant-pH MD methodology, accounting for a total of 1.25μs.

The most significant effect caused by the change to pH 7 is a global stabilization of the protein structure. We observed that some conformational transitions induced by pH 2 were possible to be reverted in many of our simulations, but only in those started from the early moments of the misfolding transition. In other words, if we stop the misfolding process before a major conformational transition takes place, we can revert it. It was not

possible to observe a complete reversibility event from a misfolded conformation. Nevertheless, we can not conclude that the transition is irreversible because we can only sample reversible phenomena that happen at sub μs timescale.

Keywords - prion protein, misfolding, pH, Constant-pH Molecular Dynamics, N-O contacts, principal component analysis

Resumo

A proteína priónica (PrP), cuja função é ainda desconhecida, é o agente patogénico responsável por um grupo de doenças conhecidas como Encefalopatias Espongiformes Transmissíveis. Neste grupo de doenças estão incluídas patologias como a doença de Creutzfeld-Jakob e a encefalopatia espongiforme bovina. A PrP na sua forma celular (PrP^C) é globular, solúvel e rica em hélices α. Esta proteína é glicosilada e encontra-se ligada covalentemente à membrana de neurónios através de uma molécula de glicosilfosfatidilinositol. A PrP^C pode sofrer $misfolding$, originando assim a forma patogénica desta proteína (PrP^{Sc}). Esta última é rica em estruturas β, podendo agregar e formar fibras amilóides, levando a uma situação patológica. O mecanismo de conversão entre a PrP^C e a PrP^{Sc} é desconhecido, bem como o seu detalhe estrutural. Pensa-se que esta transição pode ser induzida quer por uma molécula pré-formada de PrP^{Sc} quer por factores externos. A indução do $misfolding$ pelo pH foi já observada experimentalmente e pensa-se que biologicamente pode ocorrer nos endossomas (a acumulação da PrP^{Sc} nos endossomas já foi observada). O efeito do pH na PrP tem vindo a ser estudado nos últimos anos. Em particular, este efeito foi estudado, recentemente, no Grupo de Simulação Molecular do Institudo de Tecnologia Química e Biológica (ITQB) [**1**], através da utilização da técnica de dinâmica molecular a pH constante. Neste estudo observou-se que a quantidade de resíduos de aminoácidos em hélice e em beta depende fortemente do pH. O número de resíduos em hélice diminui com a diminuição do pH e o oposto acontece com o número de resíduos em estruturas do tipo β. Para além disto, foi observada numa simulação a pH 2 uma transição de

misfolding muito evidente.

O objectivo deste trabalho foi, então, estudar a reversibilidade do *misfolding* da PrP através da realização de simulações a pH 7 partindo de diferentes conformações retiradas do processo de *misfolding* observado na referida simulação a pH 2. Para tal, foram realizadas 47 simulações, perfazendo um total de 1.25μs. Nestas simulações foram utilizadas duas abordagens: procurou-se um fenómeno de reversibilidade completa através da realização de simulações longas; e estudou-se o efeito inicial após a mudança para o pH 7 através de simulações curtas.

Os resultados obtidos foram analisados através de diversas ferramentas / metodologias. A análise da estrutura secundária revelou uma grande perda de helicidade, bem como a presença de uma grande quantidade de estruturas β meta-estáveis, ou seja, estruturas β que são formadas transientemente em várias zonas da proteína. No entanto, a perda de helicidade observada a pH 7 pareceu-nos anormal e levou-nos a estudar a estrutura secundária através dos contactos entre átomos de azoto e oxigénio da cadeia principal. Este estudo revelou que os contactos de curto alcance são mais estáveis do que as hélices observadas anteriormente. Para além disso, observou-se que os contactos de longo alcance estão presentes numa quantidade elevada e apresentam também a meta-estabilidade que fora observada nas estruturas β. Estas medidas da estrutura secundária não revelaram nenhum fenómeno de reversibilidade evidente, podendo isto dever-se ao facto destas análises reportarem apenas detalhes estruturais locais. Assim sendo, foram utilizados outros tipos de análise, tendo a observação do *fold* global da proteína como o seu objectivo.

Para as simulações longas realizaram-se histogramas de raio de giração e de área acessível ao solvente. Estes estudos permitiram caracterizar alguns aspectos estruturais do PrP. Observou-se que as duas propriedades referidas apresentam uma forte correlação na maioria das simulações realizadas. No entanto, nas simulações que foram iniciadas em estados mais avançados do processo de *misfolding*, estas propriedades perdem a correlação. Este fenó-

meno pode dever-se ao facto da proteína poder já ter perdido grande parte dos contactos estruturais presentes na estrutura original. Assim, mesmo que o raio de giração tenha diminuído, resultado de uma compactação presente num novo fold, a área acessível ao solvente pode manter-se constante (ou até aumentar) uma vez que já haviam sido desfeitos uma grande quantidade de contactos (hidrófobos e pontes de hidrogénio), criando locais novos onde o solvente pode interagir com a proteína.

Os estudos de RMSD (Root Mean Square Deviation) foram realizados para todas a simulações curtas, uma vez que esta análise se revelou pouco útil para as simulações longas. Nestas análises foram evidenciados vários fenómenos: na maioria dos casos a mudança para pH 7 pareceu estabilizar a proteína; observou-se em alguns casos a proteína a dirigir-se conformacionalmente numa direcção contrária ao $fold$ típico do pH 7; no entanto, num número significativo de simulações, ocorreu o oposto e a proteína dirigiu-se para o $fold$ típico do pH 7. Através da análise por RMSD, a reversibilidade apresentou-se como um fenómeno possível, mas apenas nos momentos iniciais do processo de misfolding.

Finalmente, foi realizada uma análise de componentes principais. Dado que este tipo de análise necessita de simulações equilibradas, foram utilizados três conjuntos de simulações: as simulações longas, realizadas a pH 7 neste trabalho e as simulações a pH 2 e pH 7 realizadas anteriormente no grupo de Simulação Molecular do ITQB. Após a obtenção do novo espaço conformacional, é possível obter duas paisagens de energia utilizando os dois primeiros componentes principais – uma a pH 7 e outra a pH 2. As paisagens de energia obtidas revelaram uma boa separação entre as simulações anteriormente realizadas a pH 7 e as restantes. Assim sendo, constituem uma boa forma de avaliar fenómenos de reversibilidade que possam ter ocorrido. As simulações curtas foram então projectadas neste espaço. A análise dos resultados obtidos com o procedimento descrito foi concordante com a anterior realizada para o RMSD. A maioria das simulações moveu-se pouco na paisagem de energia, revelando um efeito estabilizador do pH 7. Algumas simulações espalharam-se na paisagem de energia e/ou na direcção

oposta à do maior poço característico do $fold$ do pH 7. No entanto, um número significativo de simulações dirigiu-se na direcção deste poço. Assim sendo, a análise de componentes principais confirmou também que o fenómeno de reversibilidade é possível, embora limitado apenas para estados ainda iniciais do processo de $misfolding$. Numa réplica em particular, foi possível observar visualmente a evolução da conformação da proteína em que esta se dirigia na direcção da estrutura original do PrP. De uma forma geral, os resultados obtidos com a análise de componentes principais está de acordo com o estudo do RMSD, ou seja, as simulações em que se observou o fenómeno da reversibilidade são as mesmas em ambos os estudos.

Como foi referido, a grande perda de helicidade observada levou-nos a comparar o campo de forças utilizado (GROMOS 53a6) com o seu antecessor (GROMOS 43a1). Este estudo revelou que, em simulações longas (e apenas nestas), o campo de forças GROMOS 53a6 parece destabilizar as hélices podendo levar a uma perda exagerada de estrutura.

A questão da reversibilidade do misfolding da PrP não foi completamente esclarecida neste estudo. De facto, não nos foi possível observar uma transição conformacional completa da forma $misfolded$ para o $fold$ original do PrP. Isto pode dever-se principalmente a dois factores: por um lado o campo de forças pode estar a prejudicar-nos nessa observação de reversibilidade; por outro lado, os tempos de simulação necessários para observar um fenómeno deste tipo podem ser demasiado grandes para serem acessíveis computacionalmente usando a nossa metodologia. No entanto, as simulações curtas realizadas parecem evitar o possível efeito negativo do campo de forças e nalgumas das simulações iniciais foi possível observar o fenómeno de reversibilidade. Numa das simulações realizadas observou-se um grande movimento de uma das hélices em direcção à sua posição original na estrutura de referência (estrutura de NMR da PrP). O movimento desta hélice já havia sido proposto como sendo uma consequência da protonação de alguns resíduos ou de mutações na PrP associadas à doença de Creutzfeld-Jacob [**2**].

Palavras-chave - proteína priónica, *misfolding*, pH, Dinâmica Molecular a pH constante, contactos N-O, análise de componentes principais

Acknowledgments

In the first place, I would like to acknowledge MM/MD and CE methodologies for being so challenging and captivating.

I would like to thank both groups and respective host institutions that accepted me: Molecular Simulation Group at ITQB and to Theoretical and Inorganic Chemistry Group at FCUL.

My acknowledgement goes also to one of my supervisors, Dr. Miguel Machuqueiro. He was a tremendous help in this work and really had lots of patience teaching me all the necessary subjects. I also thank him for all the times he had to get up from his desk to look at a image or a diagram in my screen. I would also like to thank my other supervisor, Dr. António Baptista, for all the useful discussions and for his desire to be sure that I understood all the aspects properly. To Dr. Sara Campos, I thank for all the discussions and for helping me to learn PCA and all her long scripts and programs.

I also thank Prof. Maria José Calhorda for the opportune questions and discussions and for helping me in many situations in this last year.

A special acknowledgment goes to Sofia Carvalho for saying so many times: "Yes Diogo, I understood your explanation about constant-pH MD method" and, of course, for all the emotional support that she gave me along this last year. I would also want to dedicate a special acknowledgment to my parents for asking me so few times: "Your thesis is about what?" and for keep supporting me all these years.

Finally, I thank all my friends and family for trying to understand what I really do when I go to the university.

Contents

Abstract i

Resumo iii

Acknowledgments ix

List of Figures xv

List of Tables xvi

List of Abbreviations xvii

1 Introduction 1

1.1 Protein Structure 1

1.1.1 Folded and Unfolded States 2

1.1.2 Misfolded State and Disease 3

1.2 Prion Protein 4

1.2.1 Biological Relevance – Disease 5

1.2.2 Folding – Misfolding transitions 5

1.2.3 Aim of this Work 7

1.3 Molecular Mechanics / Molecular Dynamics 9

1.4 Continuum Electrostatics 9

1.5 Constant-pH MD 10

2 Theory and Methods 13

2.1 Molecular Mechanics / Molecular Dynamics 14

2.1.1 Potential Energy Function 14
2.1.2 Force Field 19
2.1.3 Energy Minimization 20
2.1.4 Molecular Dynamics 22
2.1.5 Periodic Boundary Conditions 24
2.1.6 Non-bonded Interactions 25
2.1.7 Temperature / Pressure 26
2.1.8 Constraints 28
2.2 Continuum Electrostatics 29
2.3 Monte Carlo 34
2.4 Constant-pH MD 34
2.5 Simulations - Setup and Parameters 36
2.6 Analysis 41
2.6.1 Secondary Structure - DSSP 41
2.6.2 Secondary Structure - N-O Contacts Tool 42
2.6.3 Solvent Accessible Surface (SAS) 43
2.6.4 Radius of Gyration 43
2.6.5 Root Mean Square Deviation (RMSD) 44
2.6.6 Principal Component Analysis 45

3 Results and Discussion 47
3.1 Secondary Structure - DSSP 48
3.2 Secondary Structure - N-O Contacts 51
3.3 Radius of Gyration and SAS 58
3.4 RMSD 60
3.5 PCA 66
3.6 Force Field Comparison - 53a6 vs 43a1 79

4 Concluding Remarks 83

A Number of important contacts maintained in short simulations 89

B Histograms of Radius of Gyration and SAS of Long Simulations 97

C RMSD plots of short simulations 103

D Difference of average RMSD value vs pH2mis and pH7mis conformations in the short simulations 115

E Energy landscapes and projections of short simulations 121

Bibliography 133

List of Figures

1.1 NMR structures of huPrP^{C} 4
1.2 Representative snapshots of a Misfolding transition of PrP . . . 8

2.1 Interactions taken into account in MM models. 15
2.2 Example of a 1-4 exclusion. 18
2.3 Steepest Descent Algorithm. 21
2.4 Periodic Boundary Conditions. 25
2.5 LINCS 29
2.6 Representation of CE model of a His residue. 31
2.7 Thermodynamic cycle 32
2.8 Scheme of the Stochastic Constant-pH MD algorithm 35
2.9 Secondary Structure map of a misfolding transition in PrP . . . 37
2.10 Starting Snapshots for short simulations. 38
2.11 Starting Snapshots for long simulations. 39
2.12 Distribution of all N-O distances lower than 5.0 Å 42
2.13 Scheme of the algorithm for calculating SAS 43

3.1 Variation of Helix content in short simulations. 49
3.2 Secondary Structure map of a long simulation (starting time: 30.0 ns). 50
3.3 Ratio of helix content in pH2mis, pH7mis and long simulations per residue. 51
3.4 Comparison between variation of secondary structure according to the DSSP criterion and N-O contacts tool in a long simulation (29.0 ns) 53

3.5 Structural detail of two short range contacts 54
3.6 Variation of the number of important contacts maintained in short simulations. 55
3.7 Venn Diagrams 56
3.8 Structural detail of the three contacts present in pH2mis, pH7mis and long simulations 57
3.9 Structural detail of the zones involved in the six contacts present in pH7mis and long simulations 58
3.10 Histograms of Radius of Gyration and SAS of a long simulation. 59
3.17 Difference of average RMSD value vs pH2mis and pH7mis conformations in the short simulations 65
3.18 Energy landscapes at pH 7 using the first two PC's. 67
3.19 Energy landscapes at pH 7 and pH 2 using the first two PC's. . 68
3.25 Simulation: 8.0b. Energy landscapes at pH 2 and pH 7 using the first two PC's and projections of short simulations 75
3.26 Conformational transition observed after changing the pH to 7 in a short simulation 77

List of Tables

3.1 Average values of helix and β-structure content in the equilibrated segment (20 - 40 ns) of long simulations 80
3.2 Variation of helix and β-structure content after the first 10 ns of long simulations 81

4.1 Summary table of the results obtained with the RMSD against a central structure (RMSD), difference of RMSD average against pH2mis and pH7mis simulations (RMSD - dif) and PCA. 85

List of Abbreviations

aa	amino acid
Asp	aspartic acid
BFGS	Broyden–Fletcher–Goldfarb–Shanno
BSE	bovine spongiform encephalopathies
CE	continuum electrostatics
CJD	Creutzfeld Jacob Disease
Cter	C-terminus
FTIR	fourier transform infrared spectroscopy
GRF	Generalized Reaction Field
GSS	gerstmann straussler syndrome
Glu	glutamic acid
His	histidine
huPrP	human prion protein
ITQB	Instituto de Tecnologia Química e Biológica
l-BFGS	limited memory - Broyden–Fletcher–Goldfarb–Shanno
LPBE	linearized poisson-boltzmann equation
MC	monte carlo
MD	molecular dynamics
MM	molecular mechanics
NMR	nuclear magnetic resonance
Nter	N-terminus
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis

PC's	Principal Components
PEF	potential energy function
PrP	prion protein
PrP^C	cellular form of prion protein
PrP^{Sc}	scrapie form of prion protein
TSE	transmissible spongiform encephalopathies

Chapter 1

Introduction

1.1 Protein Structure

> "Proteins are the most versatile macromolecules of the cell [**3**]."

The spacial arrangement of atoms in a protein is called its structure [**4**]. Since 1958 we have access to protein structures [**5**] and it is possible to study proteins at the atomic level. Nowadays we have access to about 65 thousands protein structures deposited in the Protein Data Bank [**6**]. Proteins have four levels of structure:

- Primary - amino acid sequence
- Secondary - specific structures formed through regular hydrogen interactions in specific angles
- Tertiary - association of secondary structures into a specific fold
- Quaternary - association of two ore more folded chains

Although there are so many structures available, there is a high similarity between most of them. In fact, until recently, only 1195 folds were identified according to the SCOP classification [**7**].

1.1.1 Folded and Unfolded States

> "The folded state of a protein is formed by the condensation of the various secondary elements and it is stabilized by a large number of weak interactions. (Adapted from ref. [**3**])."

There are several definitions of the folded state of a protein and we decided to present one of them. However, it is necessary to clarify that folded state and native state are not synonyms. In one hand, folded state implies secondary structure and several weak interactions stabilizing a compact state. On the other hand, the native state can be a disorganized state, without defined secondary structure. Usually, in these cases, the native state acquire a defined structure when it is performing its biological function.

The unfolded state of a protein is an unstructured state without defined secondary, tertiary or quaternary structure elements. This state can be induced *in vitro* by using drastic conditions like high urea or guanidinium concentrations, high temperatures, drastic pH values etc.. The unfolded state can be seen as a mixture of several states that can be interconverted.

Except for some special cases, the folded state is the most stable one. There are two possible explanations for this stability: the thermodynamic and the kinetic hypothesis [**8**].

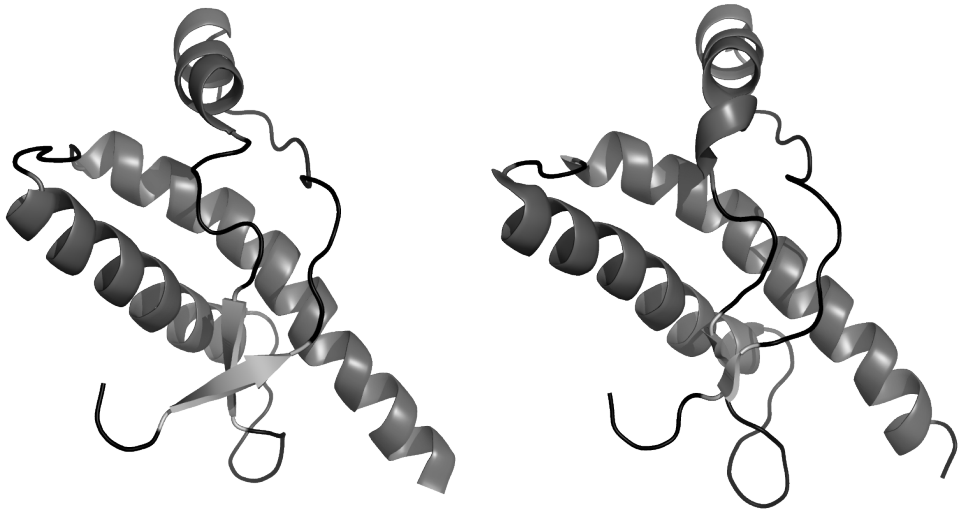
In the thermodynamic hypothesis there are several factors contributing to the variation of free energy in protein folding. Some favor the folded state (internal interactions and hydrophobic effect) and other favor the unfolded state (conformational entropy) [**9**]. Conformational entropy concerns the

global entropy of the polypeptide chain and, because the folded state is much more organized than the unfolded one, this entropic factor favors the unfolded state. However, entropy can have an opposite contribution favoring the folded state, which is the hydrophobic effect. The origin of this entropic factor is on the unfavorable restrictions imposed by the hydrophobic amino acids exposed to solvent. Finally, there is an enthalpic factor favoring the folded state which is the internal covalent and noncovalent interactions that stabilize the protein folded state: disulfide bridges, charge-charge interactions, hydrogen bonding and van der Waals interactions [**10**]. The thermodynamic hypothesis states that the folded state is thermodynamically more stable than the unfolded one [**11**]. Moreover, it was already experimentally showed that the free energy difference between the two states is small.

There is another possible explanation for the protein stability which is the kinetic stability hypothesis [**8**]. This hypothesis states that in case the unfolded state is thermodynamically more stable than the folded one, there is a high energy barrier separating the two states trapping proteins in the folded state.

1.1.2 Misfolded State and Disease

The misfolded state of a protein is a particular case of the folded state. It has both secondary and tertiary structure. Nevertheless, it differs from the native state, being a wrongly folded protein. The misfolded state can be induced by mutations or pH [**10**]. The misfolding of proteins can lead to deposit formation in the cells and, consequently, to disease. This type of diseases are generally called amyloid diseases, and Alzheimer [**12**], Parkinson [**12**] or Prion Diseases [**13**] are examples of it.



(a) PDB ID: 1QM0 **(b)** PDB ID: 1HJM

Figure 1.1: NMR structures of huPrP^{C} available in PDB corresponding to the segment 125-228. Images rendered with PyMOL software [**18**]

1.2 Prion Protein

The prion protein was discover by Prusiner in 1982 [**14**]. The cellular form of prion protein (PrP^{C}) is a glycosilated, glycosylphosphatidylinositol-anchored component of the outer surface of neuron cells and appears to have an important role in the signal transduction pathway involving Cu^{2+} [**15, 16**]. Although there are many other processes in which PrP^{C} appears to have an important role, its function remains unknown [**17**].

There are two NMR structures of human PrP (huPrP) in Protein Data Bank: 1QM0 [**19**] (**Figure 1.1a**) and 1HJM [**20**] (**Figure 1.1b**). The huPrP contains a Nter (N-terminus) region with no defined structure (aa 23-124) and a globular core (aa 125-228) composed of three α-helices (HA: aa 144-156; HB: aa 174-194; HC: aa 200-228) and a small β-sheet (b1: aa 128-131 + b2: aa 161-164) [**20**]. It has a disulfide bond between C179 and C214 and two N-glycosylation sites (N181 and N197). The Nter region is characterized by octarepeats (aa 51-91) that appear to acquire structure in

presence of Cu^{2+} or other metal [**15, 21**].

1.2.1 Biological Relevance – Disease

As previously explained, the function of PrP remains unknown. However, it is known to have a crucial role in diseases that are known as Transmissible Spongiform Encephalopathies (TSEs) [**13, 22–29**]. TSEs is a class of neurodegenerative diseases, which are fatal and for which there is no known treatment. Scrapie in sheep, Bovine Spongiform Encephalopathy (BSE) in cattle and, in humans, Creutzfeld Jacob Disease (CJD), Gerstmann Straussler Scheinker syndrome (GSS), kuru and fatal familial insomnia are examples of TSEs. Humans can acquire CJD by infection or it can be sporadic, familial or iatrogenic. The biochemical features associated with TSEs are the accumulation of amyloid deposits, vacuolation and astroglial proliferation in the brain.

1.2.2 Folding – Misfolding transitions

In 1982 Prusiner proposed that a proteinaceous infectious particle cause scrapie [**14**]. Nowadays, we know that this proteinaceous is the scrapie form of PrP (PrP^{Sc}). There are large structural differences between PrP^{C} and PrP^{Sc}, mainly in terms of secondary structure. On one hand, PrP^{C} has a high α-helix content ($\sim$42%) and a low content in β-sheet ($\sim$3%), on the other hand, PrP^{Sc} acquire a high content in β-sheet ($\sim$43%) and its α-helix content decrease to $\sim$30% [**30**] (this results were obtained by FTIR). It was observed that the conversion between PrP^{C} and PrP^{Sc} can be induced by pH [**20**], in particular Arnold *et al* in 1995 observed that PrP^{Sc} accumulates in late endossomes [**31**]. It is not known if this conversion is reversible or not, however it has been shown that, once the disulfide bond is reduced, it is possible to follow this reversibility by circular dichroism [**32**]. Gerber *et al* in 2008 also tried to observe this reversibility but were not successful because the global structural scaffold of PrP was irreversibly lost at low

pH [**33**]. Interestingly, it has been shown that it is possible to revert the unfolding of PrP induced by high concentrations of guanidine hydrochloride [**34**].

The misfolding conversion of PrP has been studied in last decade by Molecular Dynamics (MD). In 2001, Alonso *et al* observed an increase in β-sheet content performing MD simulations of 10 ns with Asp, Glu and His residues all protonated [**35**]. The temperature and mutation (Asp178Asn) effects were studied and was also observed an increase in β-sheet content [**36**]. In 2007, DeMarco *et al* observed large structural modifications and an increase in β-sheet content by protonating the Asp, Glu and His residues [**2**]. A study on the effect of protonation of His residues revealed that this event is crucial to induce the misfolding transition [**37**].

There are several studies about the misfolding conversion of PrP, however the molecular detail of the structural transition leading to the PrP^{Sc} form remains unknown. Also, the molecular structure of this misfolded protein has not been unraveled yet, mainly because this form is meta-stable and easily aggregates and precipitates, which hinders dramatically the use of NMR and x-ray diffraction. Taking in consideration these limitations, computational methods can be very useful to elucidate details of the conformational transition. Recently, in Molecular Simulation Group at ITQB (Institudo de Tecnologia Química e Biológica), this transition was studied by a Constant-pH MD method (the details of this method are briefly explained in section 2.4) [**1**]. Several simulations were performed at pH 2, 4, 5, 6 and 7 [1]. It was observed a strong pH dependence in β-sheet and α-helix content of PrP. The α-helix content decreased while the β-sheet increased by lowering the pH, in agreement with experimental studies [**30**]. Several simulations at each pH were performed and the trend is evident for almost all replicates. In one particular replicate, at pH 2, a radical transition occurred from one

[1]The simulations were started from four different structures. These structures were built from three PDB structures: 1QM0 [**19**], 1HJM [**20**] and 2PRP [**38**](NMR structure). Two structures of the N-terminal of 2PRP (structures 6 and 9) were added to 1QM0 and 1HJM. These structures were used to simulate at each pH with 3 replicates per starting structure [**1**].

conformation with high α-helix content and low β-sheet content to another one with high β-sheet content and low α-helix content which is typical of the misfolded state (**Figure 1.2**). At higher pH values (6 and 7), helices remained very stable.

1.2.3 Aim of this Work

The main goal of this project was to study the effects of reverting the pH from pH 2 to pH 7 on the conformational space of PrP. This study was focused on one replicate at pH 2 from the previous work in Molecular Simulation Group at ITQB (ref. [**1**] and section 1.2.2) which is the replicate with a sharper secondary structure transition leading to a misfolded conformation. It is not known if this conversion is reversible or not: it has been shown that, once the disulfide bond is reduced, it is possible to revert this transition [**32**]; Gerber *et al* in 2008 also tried to observe it but were not successful [**33**]; interestingly, it was possible to revert the unfolding of PrP induced by high concentrations of guanidine hydrochloride [**34**]. An important question arose from these low pH conformational transitions: are any of these transitions reversible by increasing the pH back to 7? In this work, we investigated which of the following hypothesis is correct:

- The conformational transition at low pH is completely reversible
- The conformational transition at low pH is completely irreversible
- The conformational transition at low pH is only reversible until a certain critical point

To test these possibilities, the Constant-pH MD methodology (see following sections) was used. The details of the simulations performed to tackle this problem are explained in section 2.5.

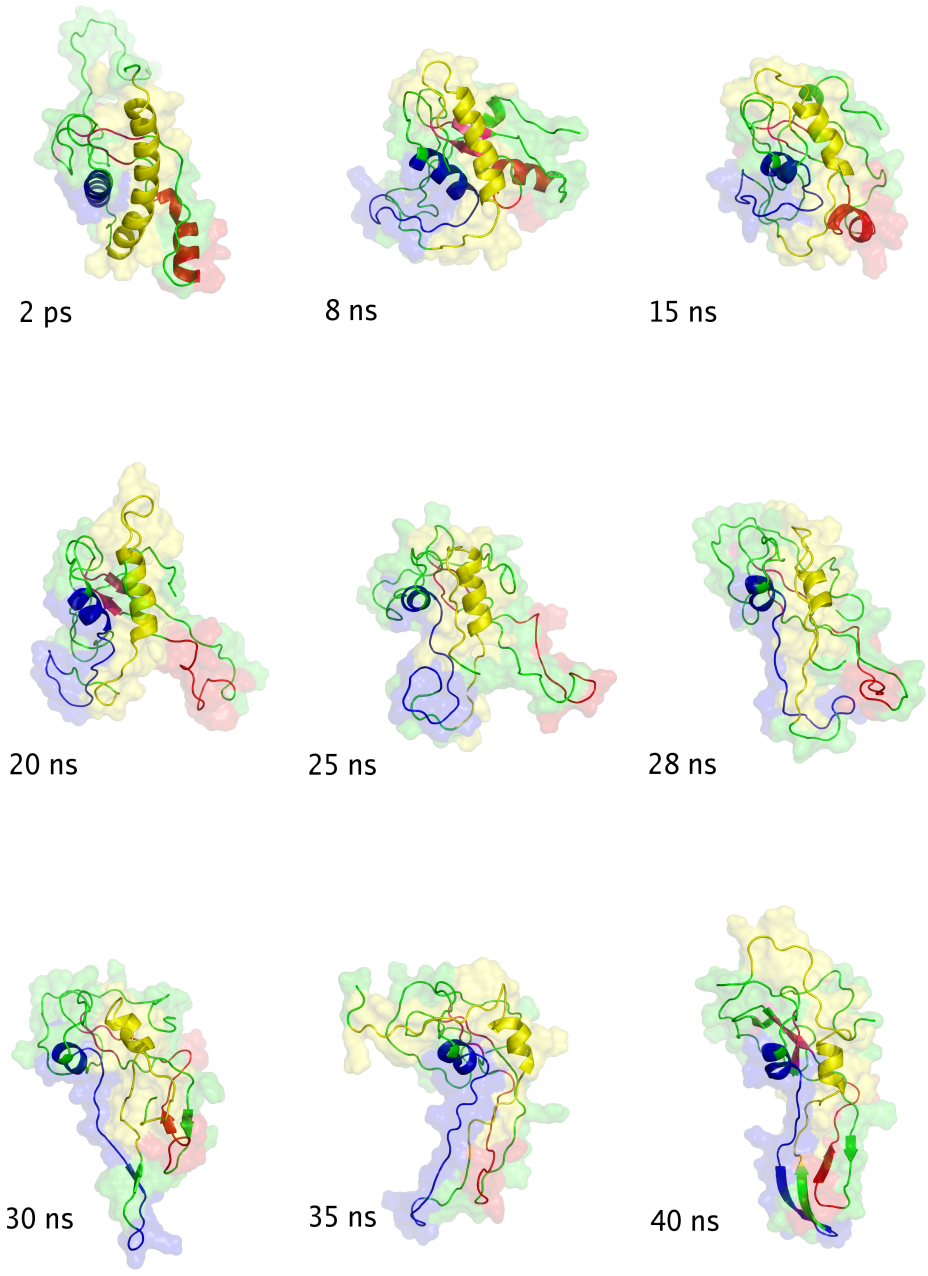


Figure 1.2: Representative snapshots of a Misfolding transition of PrP obtained at pH 2 using constant-pH MD methodology [**1**]. Image adapted with permission of Sara Campos and António Baptista.

1.3 Molecular Mechanics / Molecular Dynamics

Molecular Mechanics / Molecular Dynamics (MM/MD) methods are used to simulate the dynamic behavior of molecules in liquid phase (details about these methods in section 2.1). One of the first models of liquids involved the analysis of the behavior of a large number of gelatin balls representing molecules [**39**]. During the following decades, the models for liquids were limited to these type of methods which have many limitations like the effect of gravity [**40**].

The first computer simulation (a Monte Carlo simulation - MM/MC) was done in 1953 at the Los Alamos National Laboratories in the United States [**41**]. The first liquid simulation by solving the equations of motion was done in 1971 with water [**42**]. Finally, the first simulation of a small protein molecule was done by McCammon in 1977 with the trypsin inhibitor [**43**].

Nowadays, with the enhancing of computational power it is possible to simulate very large systems like membranes [**44**], large proteins or, even, a complete virus [**45**].

For the present work, a great limitation of standard MM/MD methods is that it does not allow to include explicitly the pH in the simulation.

1.4 Continuum Electrostatics

The Continuum Electrostatics (CE) methods are able to describe the electrostatics of a molecule as a continuous function in space (for details see section 2.2 and 2.3). There are several ways to deal with this problem and many approaches have been developed [**46–49**]. In this work, we employed a particular approach which uses a Poisson-Boltzmann (PB) model [**50–53**] and the Monte Carlo (MC) method [**40, 41, 54, 55**].

These methodologies allow us to determine the protonation state of a specific conformation of a protein, but it can not explore its conformational space.

1.5 Constant-pH MD

As previously mentioned, it is not possible to include very accurately the effect of pH using MM/MD or CE methods alone. Nevertheless, taking advantage of the complementarity of both methodologies, the stochastic titration constant-pH MD method was developed by Baptista et al [**53, 56**]. A new implementation of this method is now available and has already been successfuly applied to peptides [**57–60**] and proteins [**1, 61, 62**]. This methodology was used in this work and the details are briefly described in section 2.4.

The introduction of the pH effect in a MD simulation was originally proposed by Pepita *et Hal* in 1994 [**63**] with the study of the proton exchange between acetic acid and water. In 1997, the implicit titration method for Constant-pH MD was developed [**64**] which is based on the complementarity between MM/MD and simplified models. Since 2002 other methods were developed based on the MM/MD and CE methodologies. The main difference between these methods regards the solvation models:

- Uniform-dielectric Langevin dynamics [**65**]
- Analytical continuum solvent potential [**66**]
- Generalized Born (GB) [**67**]

A different method based only in MM/MD was developed by Börjesson and Hünenberger [**68**], however its theoretical basis is problematic [**69**]. Another constant-pH method using GB implicit solvent model and Linear Response Approximation (LRA) methods was developed by Brooks et Gal

[**70, 71**]. One of the main advantages of the stochastic titration method [**53, 56**] is the fact that (unlike the others) it uses explicit solvent. This fact can be of major importance when studying conformational transitions of peptides and proteins in water.

Chapter 2

Theory and Methods

A description of the scientific problem was performed in the first chapter. Now, it is mandatory to describe the methodologies adopted to address such problem. The goals of this chapter are to describe briefly how Molecular Mechanics/Molecular Dynamics (MM/MD), Continuum Electrostatics (CE) and Constant-pH MD work and to introduce the tools that were used in the evaluation of the results. This chapter aims to help the reader to understand more easily the implications and limitations of the used methodologies.

Proteins, like all matter, are made of atoms. Their properties result from intra and intermolecular phenomena. However, these phenomena occur at the molecular level and, in many cases, are not easily accessed experimentally. Therefore, molecular modelling and simulation methods can be useful and sometimes the only way to address them.

Molecular Modelling consists in creating a description of the system in mathematical and physical terms. Molecular Simulation is a way to predict the properties of the system using the created model. The chosen model depends on how detailed you want to describe your system. For example, for a quantum level description one would need such a high detail that

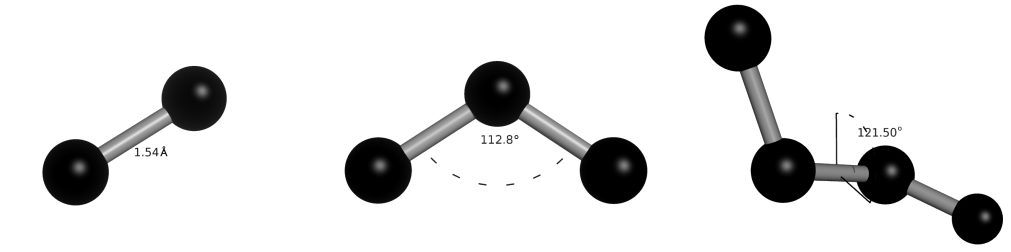
would be computationally unaffordable. In the present work, we used a classic model to describe our system (MM/MD and CE methods). Due to the high complexity of our simulations, we had to use computers with vary high calculation power.

2.1 Molecular Mechanics / Molecular Dynamics

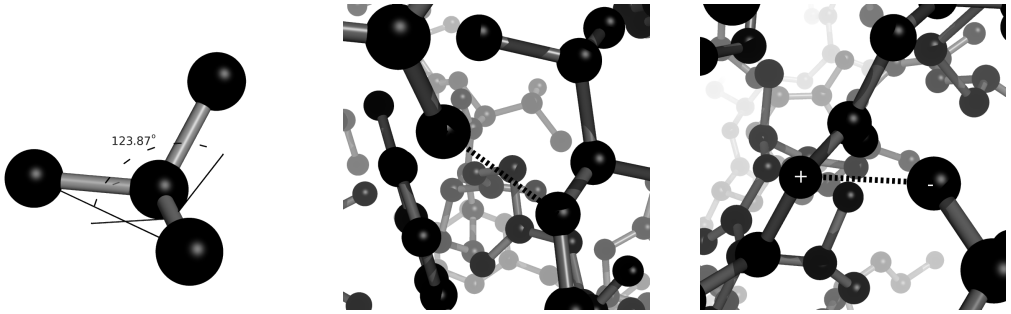
To characterize the conformational space of PrP it was needed a conformational search method and a way to model the system. We decided to use a classic model to describe it: MM / MD was used to sample the conformational space. In this type of models the Born-Oppenheimer approximation is assumed. In other words, it is assumed that the electrons can adapt to the nuclei positions very quickly so we can look only at the nuclei positions. A detailed explanation about MM/MD can be found in the following sections and complemented in references [**40**] and [**54**].

2.1.1 Potential Energy Function

The *potential energy function* (PEF) is a way to describe the protein and its interactions ignoring electrons movements and taking in consideration only the nuclei positions. The PEFs used in this project were the ones associated with the GROMOS96 43a1 [**72, 73**] and GROMOS96 53a6 [**74**] force fields. Generally, this approaches take into account four types of bonded interactions: bond length stretch (**Figure 2.1a** and **Equation 2.2**), bending angles (**Figure 2.1b** and **Equation 2.3**) and torsions of proper (**Figure 2.1c** and **Equation 2.4**) and improper (**Figure 2.1d** and **Equation 2.5**) dihedrals. The used functions also take into account two non-bonded interactions: van der Waals forces (**Figure 2.1e** and **Equation 2.6**) and electrostatic interactions (**Figure 2.1f** and **Equation 2.7**).



(a) Bond **(b)** Angle **(c)** Proper Dihedral



(d) Improper Dihedral **(e)** van der Waals force **(f)** Electrostatic Interaction

Figure 2.1: Interactions taken into account in MM models. a), b), c) and d) are bonded interactions. **e)** and **f)** are non-bonded interactions. Images rendered with PyMOL software [**18**]

The PEF takes the form presented in **Equation 2.1**.

$$V(\mathbf{r}^N) = V_b(\mathbf{r}^N) + V_a(\mathbf{r}^N) + V_{pd}(\mathbf{r}^N) + V_{id}(\mathbf{r}^N) + V_{vdw}(\mathbf{r}^N) + V_{elect}(\mathbf{r}^N) \quad (2.1)$$

$\mathbf{r}^N$ represents the vectorial positions of all atoms in the system, the first four terms correspond to bonded interactions and the last two correspond to non-bonded interactions.

The first term of the PEF ($V_b(\mathbf{r}^N)$) regards the bond length stretch and it can be described as a harmonic potential (**Equation 2.2**).

$$V_b(\mathbf{r}^N) = \sum_{n=1}^{N_b} \frac{1}{2} K_{b_n} (b_n - b_{0_n})^2 \tag{2.2}$$

K_{b_n} is the force constant between two atoms, b_n is the distance between them and b_{0_n} is the optimum distance between these atoms. This potential energy contribution is calculated for each bond in the system.

The second term of the PEF ($V_a(\mathbf{r}^N)$) addresses the angle bend. This bending between three atoms can be described as a harmonic angle potential (**Equation 2.3**).

$$V_a(\mathbf{r}^N) = \sum_{n=1}^{N_\theta} \frac{1}{2} K_{\theta_n} (\theta_n - \theta_{0_n})^2 \tag{2.3}$$

K_{θ_n} is the force constant associated with angle bend, θ_n is the angle between three atoms and θ_{0_n} is the optimum angle between these atoms. This potential energy contribution is calculated for each angle in the system.

In the third and fourth terms of PEF ($V_{pd}(\mathbf{r}^N)$ and $V_{id}(\mathbf{r}^N)$, respectively) we have the potential associated to torsion angles, proper and improper dihedral, respectively.

The proper dihedrals can be described with a periodic function with minimums and maximums at regular intervals (**Equation 2.4**).

$$V_{pd}(\mathbf{r}^N) = \sum_{n=1}^{N_\varphi} K_{\varphi_n} [1 + \cos(m_n \varphi_n - \delta_n)] \tag{2.4}$$

K_{φ_n} is the force constant associated with torsion around the dihedral angle, φ_n is the value of proper dihedral angle, m_n is the multiplicity which means the number of minima (and maxima) of energy in one complete rotation and δ_n is the reference maximum and it can be 0 or π. The potential energy of proper dihedrals is calculated to describe the torsion freedom that some dihedral angles have.

The last bonded interaction in **Equation 2.1** regards the improper dihedrals. This can also be described as a harmonic potential (**Equation 2.5**).

$$V_{id}(\mathbf{r}^N) = \sum_{n=1}^{N_\xi} \frac{1}{2} K_{\xi_n} (\xi_n - \xi_{0_n})^2 \tag{2.5}$$

K_{ξ_n} is the force constant associated with angle torsion, ξ_n is the value of improper dihedral angle between four atoms and ξ_{0_n} is the optimum angle between these atoms. This energy contribution is calculated to restrict the torsions in planar groups and in tetrahedral centers.

The last two terms of the PEF are related with non-bonded interactions: van der Waals forces ($V_{vdw}(\mathbf{r}^N)$) and electrostatic interactions ($V_{elect}(\mathbf{r}^N)$).

Van der Waals forces are described as a Lennard-Jones interaction (**Equation 2.6**).

$$V_{vdw}(\mathbf{r}^N) = \sum_{i=1}^{N} \sum_{j>i}^{N} \left(\frac{C_{12_{ij}}}{r_{ij}^{12}} - \frac{C_{6_{ij}}}{r_{ij}^{6}} \right) \tag{2.6}$$

r_{ij} is the distance between two atoms (i and j) and $C_{12_{ij}}$ and $C_{6_{ij}}$ are interaction parameters. $C_{12_{ij}}$ regards the repulsion between these two atoms while $C_{6_{ij}}$ regards attraction.

The electrostatic interactions are described according to Coulomb's law (**Equation 2.7**).

$$V_{elect}(\mathbf{r}^N) = \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{2.7}$$

r_{ij} is the distance between two atoms (i and j), q_i and q_j are the charges of these atoms, ϵ_0 is the permittivity in vacuum and ϵ_r is the relative dielectric constant.

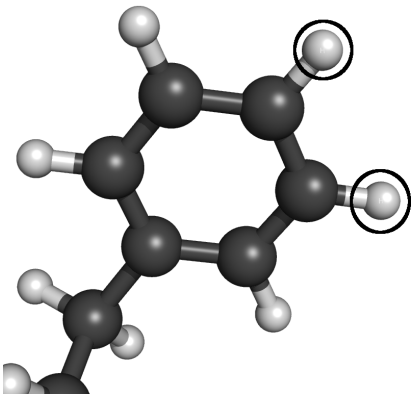


Figure 2.2: Example of a 1-4 exclusion. The hydrogens surrounded with a circle must be excluded from the non-bonded interactions to avoid the distortion in benzene plane. Image rendered with PyMOL software [**18**]

The non-bonded interactions are calculated for all pairs in the system except for those that are bonded to each other or that are separated by two covalent bonds. There is another list of exceptions, called the 1-4 exclusions, which is defined by the user. The 1-4 exclusions are used to avoid distortions in groups like benzene. An example of these exceptions is shown in **Figure 2.2**.

Force

Once $V_b(\mathbf{r}^N)$ is calculated, it is possible to compute forces on each atom (**Equation 2.8**).

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} V \tag{2.8}$$

$\nabla_{\mathbf{r}_i}$ is the gradient calculated in the position of atom i [1]. $\mathbf{F}_i$ is the resulting force on atom i. With this vectorial quantity ($\mathbf{F}_i$) it is possible to determine the pathway leading to low energy (see section 2.1.3) or it can be used to

[1] $\nabla_{\mathbf{r_i}} f = \frac{\partial f}{\partial \mathbf{r_i}} = \left(\frac{\partial f}{\partial x_i}, \frac{\partial f}{\partial y_i}, \frac{\partial f}{\partial z_i}\right)$, f is a function with three variables

calculate the acceleration, which after integration in relation to time, can be used to perform a MD simulation (see section 2.1.4).

2.1.2 Force Field

As stated in the previous section, there are several parameters necessary to describe a molecular system (K_{b_n}, b_{0_n}, K_{θ_n}, θ_n, q_i, q_j, etc.). These parameters are stored in a force field. The quality of a molecular simulation strongly depends on the quality of the used force field. From a pragmatic point of view, the force field is as good as its ability to reproduce experimental evidences. In this work, two GROMOS force fields [**75, 76**] were used: 43a1 [**72, 73**] and 53a6 [**74**]. There are many GROMOS force field variations based on the same principles[**77**]:

- They have a simple functional form and a limited set of different atom, bond, angle and dihedral angle types
- They use the united atom approach which means that non-polar hydrogens are not treated explicitly, but they are considered as a single interaction site together with the atom they are bonded to. The same approach is not used on polar hydrogens because they play a very important role in hydrogen bonding.

The GROMOS 53a6 was used because it is the most recent and improved version and it has been used previously with our implementation of the Constant-pH MD method [**1, 57, 59–62**]. It was also the force field adopted for the work in which our project is based on [**1**]. The GROMOS 43a1 was used to evaluate the different effects of both force fields when compared with each other. It has been proposed that the differences between these force fields can play an significant role in helix content [**78, 79**].

2.1.3 Energy Minimization

The PEF is a multidimensional function of the coordinates of a system and, with force field parameters, allows to calculate a potential energy value for each conformation of that system. Search for energy minima of this function is very useful in molecular simulations since they correspond to high probability states of the system and are good starting points for MD simulations.

Energy minimization is a method to search in the conformational space for an energy minimum of the system. It uses a PEF to describe that conformational space. An energy minimum is a combination of $\mathbf{r}^N$ coordinates that minimize the $V_b(\mathbf{r}^N)$ function. In other words, it is a point where the gradient of $V_b(\mathbf{r}^N)$ is zero (or does not exist) and the determinant of Hessian matrix [2] is positive. Since this function has a very large number of variables, the energy landscape is very complex and the number of minima and maxima are also very large, rendering it very difficult to find the absolute minimum. Minimization algorithms are used to search for a local minimum of energy of the system in the energy landscape.

There are several methods to perform an energy minimization. In this section, we stress the ones used in this project: steepest descent and l-bfgs. In a general way, the best solution to perform an energy minimization of complex systems is to use more than one method.

Steepest Descent

The Steepest Descent is a very simple and robust algorithm and it is easy to implement. Basically, the algorithm makes the energy to go down along the steepest descent in the energy landscape, in other words, in the direction of the maximum component of the force. **Figure 2.3** is a representation of a possible pathway undertaken by this algorithm in one dimension. After

[2] Hessian matrix is the square matrix (n × n) of second-order partial derivatives of a function with n variables

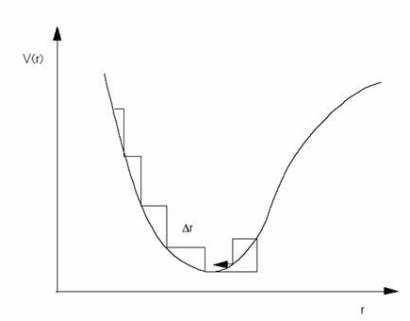


Figure 2.3: Steepest Descent Algorithm. Representation of the pathway undertaken by steepest descent algorithm in one dimension.

calculating the potential energy and the forces, the next positions of the system ($\mathbf{r}_i(n+1)$) can be obtained (**Equation 2.9**).

$$\mathbf{r}_i(n+1) = \mathbf{r}_i(n) + \frac{\mathbf{F}_i(n)}{\max\left(|\mathbf{F}_i(n)|\right)} h_n \tag{2.9}$$

$\max\left(|\mathbf{F}_i(n)|\right)$ is the highest absolute value of the components of the force. h_n is the maximum displacement and is defined by the user. After calculating the new positions, it is possible to calculate a new energy value and compute the forces. If $V(n+1) < V(n)$ the new positions are accepted and h_{n+1} becomes $1.2h_n$, if $V(n+1) > V(n)$ the new positions are rejected and h_{n+1} becomes $0.2h_n$. The algorithm stops after a number of cycles defined by the user or when $\max\left(|\mathbf{F}_i(n)|\right)$ is lower than a select value. The last conformation is not necessarily the lowest energy conformation of the system but it has a lower energy than the first one. Even though this method does not allow to jump along local minima it converges very quickly to a local minimum (or near this). It can be a very useful method to correct strange conformations and as a first step of a more complex minimization procedure.

l-BFGS

l-BFGS (limited memory - Broyden–Fletcher–Goldfarb–Shanno) [**80**] is a minimization method much more sophisticated than Steepest Descent. This

method is an improvement of the original BFGS which tries to approximate the inverse Hessian matrix but whose detailed explanation is beyond the scope of this text (see detailed information in ref. [**80**]). Basically, the l-BFGS method is a way to use the BFGS method but with lower memory cost.

The l-BFGS method allows to jump between local minima and give better results than Steepest Descent, however is much slower and should not be used as the first step of a minimization procedure. Although this method allows jumps between local minima, it remains very unlikely to reach the global minimum.

2.1.4 Molecular Dynamics

With the PEF that describes our system and the corresponding forces, we can simulate motion. Molecular Dynamics (MD) is one of the most used methods to simulate the dynamics of a protein. MD uses the Newton's equations of motion and calculates a specific trajectory which corresponds to the changes of positions and velocities of the atoms along the simulation time. With this method, it is possible to explore the conformational space of our system and predict its preferred conformations. The forces allows us to calculate the acceleration of each particle at instant t using Newton's second law (**Equation 2.10**).

$$\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F}_i}{m_i} \tag{2.10}$$

With $\mathbf{r}_i$ being the coordinates of one atom (or particle in a general way), m_i being the mass of that atom and $\mathbf{F}_i$ the force on atom i at instant t. Starting from **Equation 2.10**, it is possible to integrate it to obtain the new positions and velocities of all atoms in the system. However, the motions of all particles are coupled which makes the problem impossible to solve analytically, being necessary the use of numerical methods. Basically, these

methods integrate the equations of motion in very small steps and the force on each particle is a result of its interaction with other particles. From the force, acceleration, velocities and positions are calculated at instant t, which are finally used to calculate the new positions and velocities at instant $t + \Delta t$. This procedure is repeated many times until the simulation is finished (the number of steps is defined by the user).

One possible integration method is the leap-frog algorithm developed by Hockney in 1974 . This algorithm uses the positions $\mathbf{r}_i$ at instant t and the velocities at instant $t - \frac{\Delta t}{2}$ and calculates the new positions and velocities according with **Equation 2.11** and **Equation 2.12**.

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i \left(t + \frac{\Delta t}{2} \right) \Delta t \tag{2.11}$$

$$\mathbf{v}_i \left(t + \frac{\Delta t}{2} \right) = \mathbf{v}_i \left(t - \frac{\Delta t}{2} \right) + \frac{\mathbf{F}_i(t)}{m_i} \Delta t \tag{2.12}$$

Initiation

It was mentioned that before starting a MD simulation it is necessary to perform an energy minimization of the system. There is another essential procedure to be done after energy minimization, which is called the initiation. Initiation simulations have two objectives: generate velocities and start the systems movement without introducing too much instability that could lead to artifacts. Velocities can be generated in a very simple way: randomly. However, the magnitude of the velocity vector depends on the chosen temperature. A seed is chosen by the user and used to generate the velocities. To avoid some artifacts like loss of tertiary structure some tricks are performed in the initiation step. Normally, the initiation is started with all atoms of the protein restrained to allow water to adapt to the protein. Then the protein is released in a stepwise manner, for example, first the side chains and finally all the protein is released. There are several procedures to perform an initiation MD simulation, but the goal is the same.

Restraints are external forces applied to some (or all) atoms of the system to limit its movement. This method is used to avoid abnormal conformations or to include information from experiment. It is done by adding a new term to the PEF which is responsible to limit the movement of the desired atom with a force constant K_{pr}. This new PEF is given by **Equation 2.13**.

$$V_{pr}(\mathbf{r}_i) = \frac{1}{2} K_{pr} \left\| \mathbf{r}_i - \mathbf{R}_i \right\|^2 \tag{2.13}$$

$\mathbf{R}_i$ are the reference positions. This method is usually used to avoid drastic rearrangements due to a non-equilibrated solvent originating large forces in the protein atoms. When the protein movements seem normal, the initiation step can be stopped and production MD simulation can be started.

2.1.5 Periodic Boundary Conditions

In MM/MD systems the protein is normally surrounded by solvent molecules, however, we are limited to a finite number of molecules, which leads to a problem. This problem is solved using Periodic Boundary Conditions (PBC) (**Figure 2.4**). This method attempts to approximate the effect of a macroscopic amount of molecules and consists in surrounding our system with copies of itself which renders the system infinite. The coordinates of the copies are calculated by adding or subtracting multiples of the vectors that describe our box. Therefore, when a molecule leaves the central box from one side, it enters on the opposite side, which results in the absence of physical walls in our system. The main approximation of a system using PBC is that it simulates an infinite dilution, which is not completely coherent with physical reality.

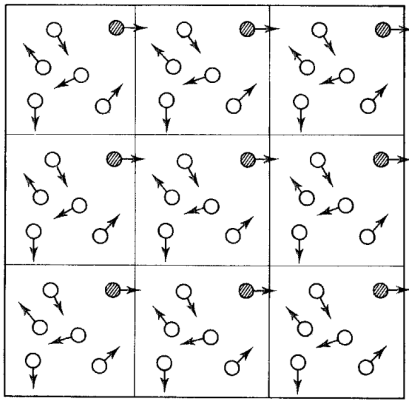


Figure 2.4: Periodic Boundary Conditions. Representation of a periodic system according with Periodic Boundary Conditions (adapted from [**54**]).

2.1.6 Non-bonded Interactions

The use of PBC creates a problem in the treatment of non-bonded interactions since we can not treat all non-bonded interactions of an infinite system in all directions of the space. There are several approaches to deal with this problem and the simplest one is to apply a cutoff radius to treat only the interactions with the atoms closer to the reference one. However, there are much more sophisticated methods with low increase computational cost. Follows a description of the methods used in this work: the twin range method to deal with van der Waals interactions [**81**] and Generalized Reaction Field method to deal with electrostatic interactions [**82**].

The twin range method [**81**] uses two cut-off radii surrounding the atom i. There are two groups contributing to long range van der Waals interaction: the neighbor list (atoms inside the shortest cut-off) and the so-called long range force (atoms between the shortest and the longest cut-offs). The neighbor list is updated in every step of the simulation and the value for the interaction is also calculated in every step. The list of atoms between the two cut-offs is determined every n steps and n is defined by the user. Once these atoms are determined, the values of interaction are calculated and stored. In a single step there are two values contributing to the van der

Waals interaction with atom i, the value of interaction with the neighbor atoms and the long range force. The atoms beyond the second cut-off value are not considered in the twin-range method.

The Generalized Reaction Field (GRF) method [**82**] is a particular case of the Reaction Field method. The main difference between them is that, in GRF, the ionic strength is considered directly as an external parameter. GRF assumes that the bulk solvent surrounding the solute is homogeneous and it can be treated as a continuous medium. This method divides the system in two parts: the inner region where the atomic charges are explicitly treated with the twin-range method (dielectric constant ϵ_{in}) and the outer region, beyond a defined distance which is treated as a continuum medium with a dielectric constant ϵ_{out} and an ionic strength I.

2.1.7 Temperature / Pressure

For a correct simulation of solvated proteins, it is necessary to take into account external factors like Temperature and Pressure. MD simulations are done at constant Temperature and Pressure in order to mimic physiological conditions. To deal with such conditions, it is necessary to use specific algorithms. The general idea of such algorithms is to correct the velocities and the positions of the atoms to reproduce the desired temperature and pressure, respectively. There are different methods to deal with pressure and temperature, however we will only describe the ones used in this work: the Berendsen bath of Temperature and Pressure [**83**].

Temperature

The temperature is a thermodynamic property given by the average kinetic energy of the system over the simulation time. The way to maintain the temperature of the system constant is to couple it to an external thermal bath. This external bath is an external source of thermal energy that can

provide or remove energy from the system. This is done by scaling the velocities of the atoms in the system. The velocity variation is proportional to the difference between the bath temperature and the system temperature (**Equation 2.14**).

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t} = \frac{T_{bath} - T(t)}{\tau} \tag{2.14}$$

τ is coupling constant which determines how the bath temperature and system temperature are coupled. With this method, the system temperature decays exponentially to the desired temperature. The temperature difference between two steps is given by **Equation 2.15**.

$$\Delta T = \frac{\Delta t}{\tau} \left(T_{bath} - T(t)\right) \tag{2.15}$$

Finally the scaling factor (λ) applied to the velocities is given by **Equation 2.16**.

$$\lambda = \left[1 + \frac{\Delta t}{\tau} \left(\frac{T_{bath}}{T(t)} - 1\right)\right]^{\frac{1}{2}} \tag{2.16}$$

The coupling strength depends on the value of τ. A high value of τ means a small temperature coupling and $vice - versa$. Usually, a lower value of τ is used in the first steps of simulation to ensure that the temperature converge quickly to the desired value.

Pressure

In the case of pressure the idea is approximately the same: coupling our system to an external bath of pressure in this case. The basic idea is that, if the system presents a higher value of pressure than the desired one, the volume of the simulation box should be increased and $vice - versa$. The pressure variation velocity is given by the **Equation 2.17**.

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = \frac{P_{bath} - P(t)}{\tau_P} \tag{2.17}$$

τ_P is the coupling constant in analogy to τ present in temperature. The volume of the simulation box is scaled by a factor μ (**Equation 2.18**).

$$\mu = 1 - \beta \frac{\Delta t}{3\tau_P} \left(P_{bath} - P(t)\right) \tag{2.18}$$

β is the isothermal compressibility of the system. The new atoms positions are given by **Equation 2.19**.

$$\mathbf{r_i'} = \mu^{\frac{1}{3}} \mathbf{r_i} \tag{2.19}$$

Once again, the higher the value of τ_P, the less Pressure is coupled to the desired value and $vice - versa$.

2.1.8 Constraints

One of the most common problems of this type of simulations is that they are very time consuming. Therefore, some approximations are commonly used in order to minimize the computational cost. One way to speed up the simulations is to use a high integration step, however it entails several risks because there are several phenomena happening at very short times, like bond stretching. One way to increase the integration step without loosing accuracy significantly is the use of constraint algorithms.

There are several methods to constraint bond and angles like LINCS [**84**] which can be used to constraint bond lengths and some angles, SHAKE [**85**] which is slower and less stable then LINCS, and SETTLE [**86**] which is an analytical solution of SHAKE specifically developed for water molecules.

The LINCS algorithm is a very useful tool to constraint bond lengths. It restores the bond lengths to the equilibrium values after an unrestricted

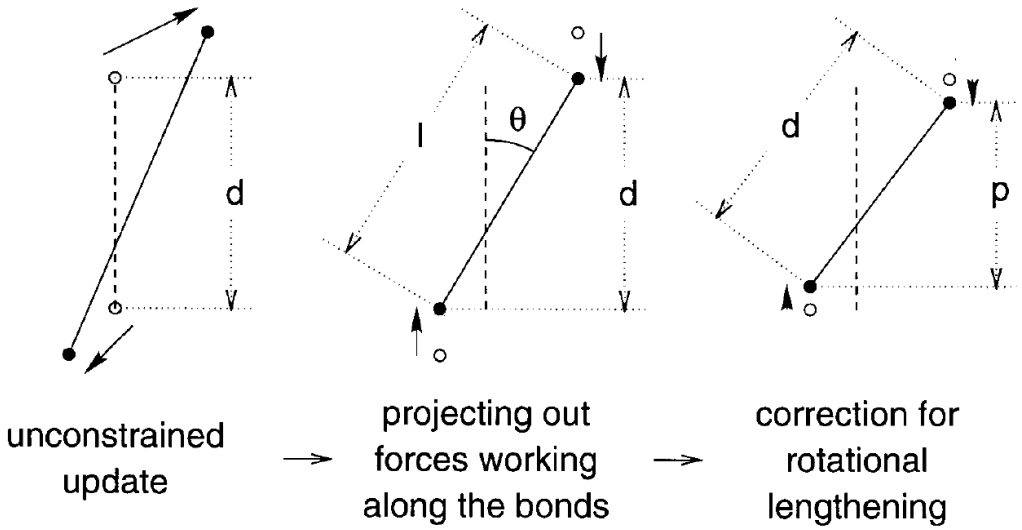


Figure 2.5: LINCS algorithm for bond length constraint (adapted from [**84**]).

integration step. The algorithm works with two steps (see **Figure 2.5**): first, the new bonds are projected in the old ones and then a correction is applied to adjust the bond length to the reference minimum.

The SETTLE algorithm is usually used to constraint bond length and angles in water molecules. Similar to LINCS, the SETTLE algorithm restores bond lengths and angles to the equilibrium values after an unrestricted integration step.

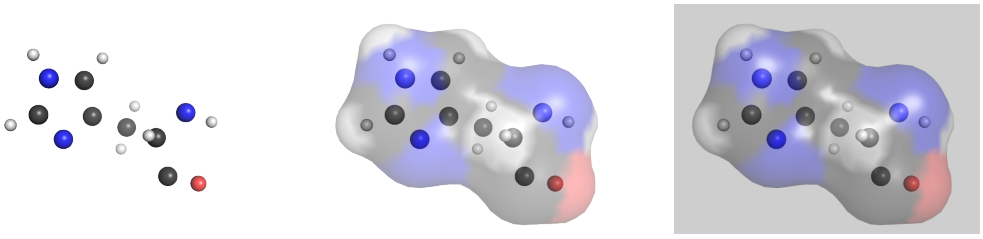
2.2 Continuum Electrostatics

As explained in the previous section (2.1), Molecular Dynamics was used to simulate the dynamic behavior of PrP. To perform a Constant-pH MD simulation, it is necessary to evaluate the protonation state of a protein conformation throughout the simulation run (more details in sections 1.5 and 2.4). The protonation state of a protein can be obtained from Continuum Electrostatics (CE) calculations.

The treatment of the electrostatic behavior of a protein is very useful in the understanding of many biochemical processes. Charged and polar groups, which are found ubiquitously in biological macromolecules, have an important role in processes like enzymatic catalysis, protein stability or biomolecular recognition [**87**]. There are several methods to deal with electrostatic interactions in proteins [**46**–**49**]. In this work, the protonation equilibria was assessed using the Poisson-Boltzmann (PB) model [**50**–**53**] in which the protein is treated as a rigid body (this is one of the approximations of these models), and the protonation states were sampled with Monte Carlo (MC) simulations [**40, 41, 54, 55**].

To perform a PB calculation it is necessary to determine the electrostatic potential of our system, which is a continuous function in space. To do that, some approximations are necessary. In PB models, the protein is described as a continuous region with a single value of dielectric constant ϵ_{in} (**Figure 2.6b**). This approximation tries to describe an instantaneous reorganization of dipoles. The same approximation is done to the solvent, which is treated implicitly with a single value of dielectric constant ϵ_{out} (**Figure 2.6c**). Normally, it is used a low value of dielectric constant for solute and a high value of dielectric constant for water (usually 80). The higher the dielectric constant the more reorganizable is the material. The dielectric constant of water is large because water can quickly adopt a conformation around a charge and can easily shield and stabilize two ions with same charge very close to each other. In the protein interior, the opposite happens and the protein can not adapt quickly to a change in the electrostatic environment, hence its dielectric constant is much lower. The PB model takes also into account the atomic charges of the solute in the calculation of the electrostatic potential (**Figure 2.6a**). Finally, the ions in solution are not treated explicitly, they are taken into account by using an ionic strength (I) (**Figure 2.6c**).

The treatment of electrostatic interactions in a protein surrounded by solvent with an ionic strength I can be done with the linearized Poisson-Boltzmann equation (LPBE) (**Equation 2.20**).



(a) His atomic charges

(b) His reorganization ϵ_{in}

(c) Water reorganization ϵ_{out} and inclusion of ionic strength I

Figure 2.6: Representation of CE model of a His residue.

$$\nabla \cdot [\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})] - \kappa^2(\mathbf{r})\epsilon(\mathbf{r})\Phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{2.20}$$

$\mathbf{r}$ is the vectorial positions of all protein atoms, $\epsilon(\mathbf{r})$ is the dielectric constant, $\Phi(\mathbf{r})$ is the electrostatic potential, $\rho(\mathbf{r})$ is the charge density and $\kappa(\mathbf{r})$ is the so-called reciprocal Debye length and is given by **Equation 2.21**. All these four functions are defined in each point of the space ($\mathbf{r}$).

$$\kappa(\mathbf{r}) = \begin{cases} \left(\dfrac{8\pi e^2 I}{\epsilon_{out} k_{\mathrm{B}} T}\right)^{1/2} & \text{if } \mathbf{r} \text{ is in an ion accessible region} \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

I is the ionic strength of the solvent, ϵ_{out} is the dielectric constant of the solvent, k_{B} is the Boltzmann constant and T is the absolute temperature. $\kappa(\mathbf{r})$ is zero when the $\mathbf{r}$ is in an inaccessible region, in our case, the interior of the protein. If I is zero, the LPBE takes the form of the Poisson equation (**Equation 2.22**).

$$\nabla \cdot [\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{2.22}$$

Equation 2.20 or **Equation 2.22** are used to estimate the electrostatic potential at different positions of our system. However, these equations

have no simple analytical solution and, therefore, can be solved on a cubic lattice using a finite difference procedure.

The electrostatic potential can be used to estimate the *protonation free energy*. In other words, it can be used to calculate the free energy difference between a reference protonation state and another state (called **a**, for example). However, this calculation is, also, not straightforward and it has been done using a thermodynamic cycle (**Figure 2.7**).

$$\begin{array}{ccc} \mathrm{A^{sol}} + \mathrm{H^+} & \xrightarrow{\Delta G^\circ{}_{\mathrm{sol}}(\mathrm{A}\to\mathrm{AH})} & \mathrm{AH^{sol}} \\ \Big\downarrow{\scriptstyle \Delta G^\circ{}_{\mathrm{sol}\to\mathrm{P}}(\mathrm{A})} & & \Big\downarrow{\scriptstyle \Delta G^\circ{}_{\mathrm{sol}\to\mathrm{P}}(\mathrm{AH})} \\ \mathrm{A^P} + \mathrm{H^+} & \xrightarrow{\Delta G^\circ{}_{\mathrm{P}}(\mathrm{A}\to\mathrm{AH})} & \mathrm{AH^P} \end{array}$$

Figure 2.7: Thermodynamic cycle involving protein and model compounds.

In **Figure 2.7**, a thermodynamic cycle for one single titrable site is represented and this formalism must be generalized for all titrable sites in the protein. In the figure, $\mathrm{A^{sol}}$ and $\mathrm{A^P}$ regard the deprotonated forms in solution and protein environment, respectively; $\mathrm{AH^{sol}}$ and $\mathrm{AH^P}$ regard the equivalent protonated forms. From the thermodynamic cycle, it is possible to write an expression with standard free energy difference of protonating the site in the protein ($\Delta G^\circ_P(A \rightarrow AH)$) as function of the other three terms of the cycle (**Equation 2.23**).

$$\begin{aligned} \Delta G^\circ{}_{\mathrm{P}}(\mathrm{A}\rightarrow\mathrm{AH}) &= \Delta G^\circ{}_{\mathrm{sol}}(\mathrm{A}\rightarrow\mathrm{AH}) + \Delta G^\circ{}_{\mathrm{sol}\to\mathrm{P}}(\mathrm{AH}) - \Delta G^\circ{}_{\mathrm{sol}\to\mathrm{P}}(\mathrm{A}) \\ &= \Delta G^\circ{}_{\mathrm{sol}}(\mathrm{A}\rightarrow\mathrm{AH}) + \Delta\Delta G^\circ{}_{\mathrm{sol}\to\mathrm{P}}(\mathrm{A}\rightarrow\mathrm{AH}) \end{aligned} \tag{2.23}$$

The Gibbs free energy of deprotonation can be converted to a $\mathrm{p}K_\mathrm{a}$ value using **Equation 2.24**.

$$\mathrm{p}K_\mathrm{a} = \frac{\Delta G^\circ}{2.3 k_\mathrm{B} T} \tag{2.24}$$

With **Equation 2.23** and **Equation 2.24** it is possible to determine the pK_a of one site P (**Equation 2.25**).

$$\begin{aligned} \mathrm{p}K_a(\mathrm{P}) &= \mathrm{p}K_a(\mathrm{sol}) + \frac{1}{2.3k_BT}\Delta\Delta G^\circ{}_{\mathrm{sol}\to\mathrm{P}}(\mathrm{AH}\to\mathrm{A}) \\ &= \mathrm{p}K_{\mathrm{int}} + \frac{1}{2.3k_BT}\Delta G^\circ{}_{\mathrm{interact}}(\mathrm{P}) \end{aligned} \tag{2.25}$$

pK_a(P) is the pK_a of site P in protein environment and pK_a(sol) is the pK_a of the titrable site in solution (the so-called pK_{mod}). The calculation of pK_a(P) can be done with pK_{mod} and the contributions of changing from solvent to protein environment. pK_{int} regards the pK_{mod} and the interaction of the titrable site with the all other charges in the protein present in residues that are not titrating (the so-called background charges) and with all other titrating sites when they are all neutral, so it is pH-independent. $\Delta G^\circ{}_{\mathrm{interact}}$(P) depends on the pH and regards the contribution of the interaction of site P with all other titrating sites in the protein. The solution of the LPBE (the electrostatic potential) is then used to calculate the terms associated with the solvation effects and the interaction between the titrable sites in the protein.

Now, we can come back to protonation state **a** which is a vector with as many terms as titrable sites in the protein and a_i is 0 if the site i is neutral and 1 if site i is charged [3]. Finally, it is possible to calculate the free energy of changing from a reference state to a determined state **a** (**Equation 2.26**).

$$\Delta G^\circ(\mathbf{a}) = -2.3k_BT\sum_i a_i\gamma_i(\mathrm{p}K_{\mathrm{int},i} - \mathrm{pH}) + \sum_i\sum_{j\neq i} a_i a_j \Delta W_{ij} \tag{2.26}$$

$\Delta G^\circ(\mathbf{a})$ is the free energy of the change, γ_i is the charge of site i when it is

[3] Since titrable groups in proteins may have alternative proton positions (tautomers) a_i can acquire different values corresponding to different tautomers. For more details see reference [**88**]

ionized and ΔW_{ij} is essentially the interaction free energy between ionized sites i and j.

2.3 Monte Carlo

With the calculated free energy of changing from a reference state to state **a**, we can estimate a protonation state for our protein in that specific conformation. However, if the protein has a large number of titrable sites, it becomes impossible to calculate all combinations. To deal with this problem, it can be used a Monte Carlo (MC) sampling method, which undergoes through all sites iteratively and evaluates protonation changes. The new protonation state is accepted if $\Delta G^{\circ}(\mathbf{a}) < 0$ or with a probability of $e^{-\Delta G^{\circ}(\mathbf{a})/RT}$ if $\Delta G^{\circ}(\mathbf{a}) > 0$ (Metropolis criterion [**41**]).

After a determined number of MC steps a correct sample of the possible states is obtained. With the obtained sampling it is possible to calculate several properties. It is possible to obtain a final protonation state from the MC run (the one used in Constant-pH MD methdology - see next section), which should be representative of the used conformation. It is also possible, to calculate the average protonation of each titrating site. It is even possible to estimate the pK_{a} for each titrable residue from its titration curve, if calculations are made at different pH values.

2.4 Constant-pH MD

Simulating pH effects on peptides or proteins is not a trivial task. The dynamic behavior can be simulated by standard MM/MD (see section 2.1) which can sample the conformational space of biomolecules with high accuracy. However, to simulate a protein at a specific pH value, it is necessary to let the protonation state of titrable sites change and standard MM/MD is not able to do such task. On the other hand, PB/MC methods (see

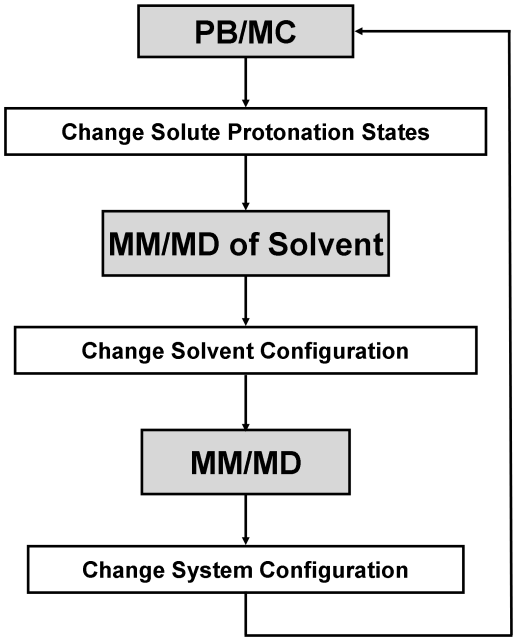


Figure 2.8: Scheme of the Stochastic Constant-pH MD algorithm (adapted from [**56**]).

sections 2.2 and 2.3) can be used to sample the protonation states in different conformations of a protein, but cannot take explicitly into account the conformational variability in each calculation. Taking into account the complementarity of standard MM/MD and PB/MC methods, Baptista et al [**53, 56**] developed the stochastic constant-pH MD method [4]. The overall scheme of this Constant-pH MD method is represented in **Figure 2.8**.

The method works in a cycle with three main steps, with the loop running until the end of the desired simulation time:

- **PB/MC** - The first step is a PB/MC calculation to sample a suitable protonation state for the first conformation of the system.
- **MM/MD of Solvent** - A short MM/MD segment is performed with frozen solute to let the water molecules adapt to the new protonation

[4]For an overall view of other constant-pH MD methods see section 1.5

state. This step allows for the solvent to relax to the newly presented charges coming from PB/MC calculations.

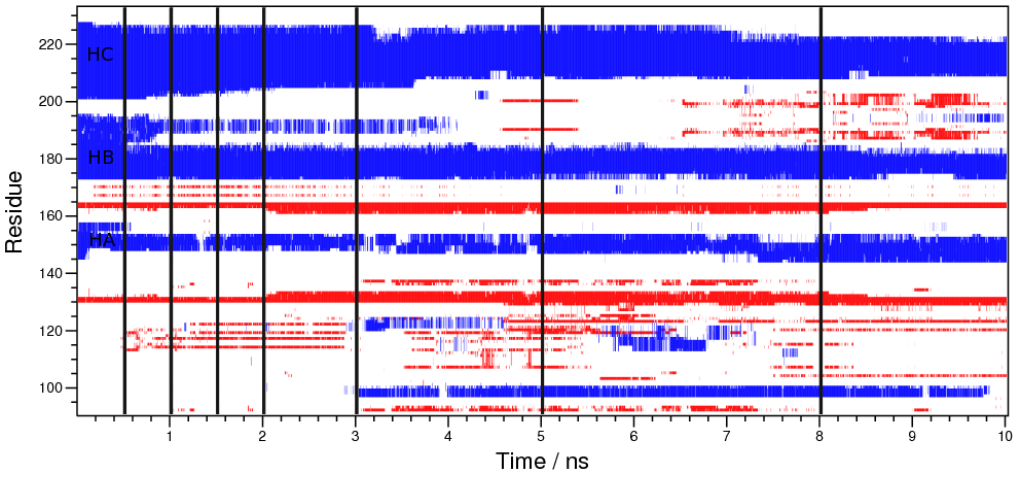
- **MM/MD** - The final MM/MD step is the production run with the system unfrozen. The last conformation is then used as input to the first step of the next cycle.

2.5 Simulations - Setup and Parameters

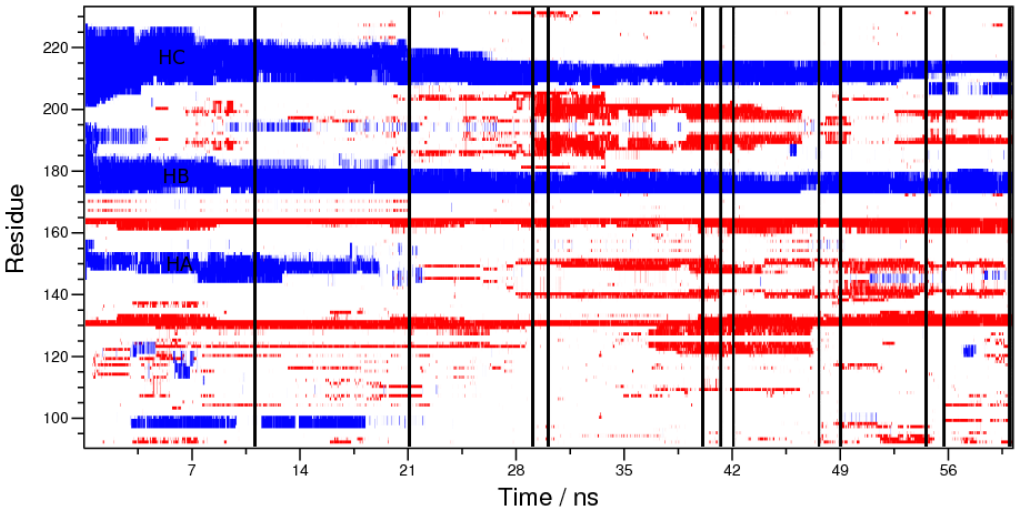
In order to study the effects of changing the pH from 2 to 7 during a misfolding transition, we used one particular simulation obtained in Molecular Simulation group at ITQB [**1**]. The observed transition is represented in **Figure 1.2** and **2.9**.

This particular 60 ns simulation originated a very pronounced transition between a very stable 'well-folded' structure at pH 7 to a myriad of meta-stable conformations at pH 2. **Figure 2.9** shows that the helix content decreases very quickly in the first nanoseconds of simulation. The HA is completely lost at $t \simeq 20$ns and the size of HB and HC is drastically reduced in the first 27 ns of simulation. On the other hand, the β-sheet natively present in PrP^C are maintained during the simulation time and, in fact, new β-sheet content is formed at $t \simeq 30$ns. This newly formed β structures are meta-stable since they are not always present. The secondary structure map indicates that PrP acquires more conformational freedom at pH 2.

In this work, a total number of 47 simulations were performed with a total time of 1.25μs. The simulations were started from different points of the previous simulation at pH 2. 21 short simulations (10 ns) (**Figure 2.10**) starting from different times were performed: 0.5 ns, 1.0 ns, 1.5 ns, 2.0 ns, 3.0 ns, 5.0 ns and 8.0 ns; three different replicates per each starting point, denoted, for example, as 0.5a, 0.5b, 0.5c, etc. Also, 12 long simulations (40 ns) were performed (**Figure 2.11**) started from: 11.0 ns, 21.0 ns, 29.0 ns, 30.0 ns, 40.0 ns, 41.2 ns, 42.0 ns, 47.5 ns, 49.0 ns, 54.5 ns, 55.7 ns and

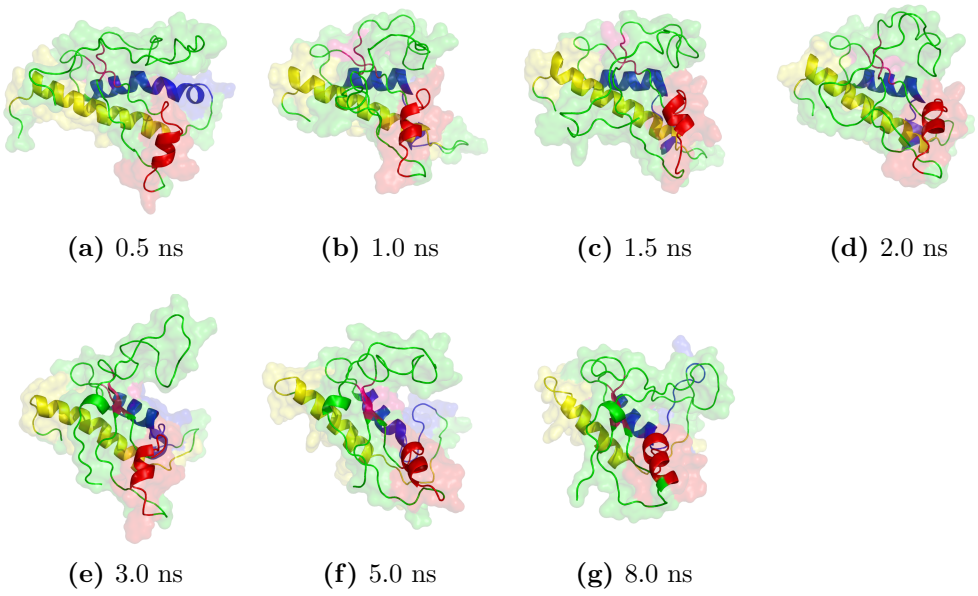


(a) First 10 ns - Starting points for short simulations



(b) All simulation (60 ns) - Starting points for long simulations

Figure 2.9: Secondary Structure map of a misfolding transition in PrP obtained at pH 2 using the constant-pH MD methodology [**1**]. ■ represents residues in β-sheet or β-bridge ■ represents the residues in helices. The vertical black lines are the simulation times from which we started our simulations. Images adapted with permission of Sara Campos and António Baptista.



(a) 0.5 ns **(b)** 1.0 ns **(c)** 1.5 ns **(d)** 2.0 ns

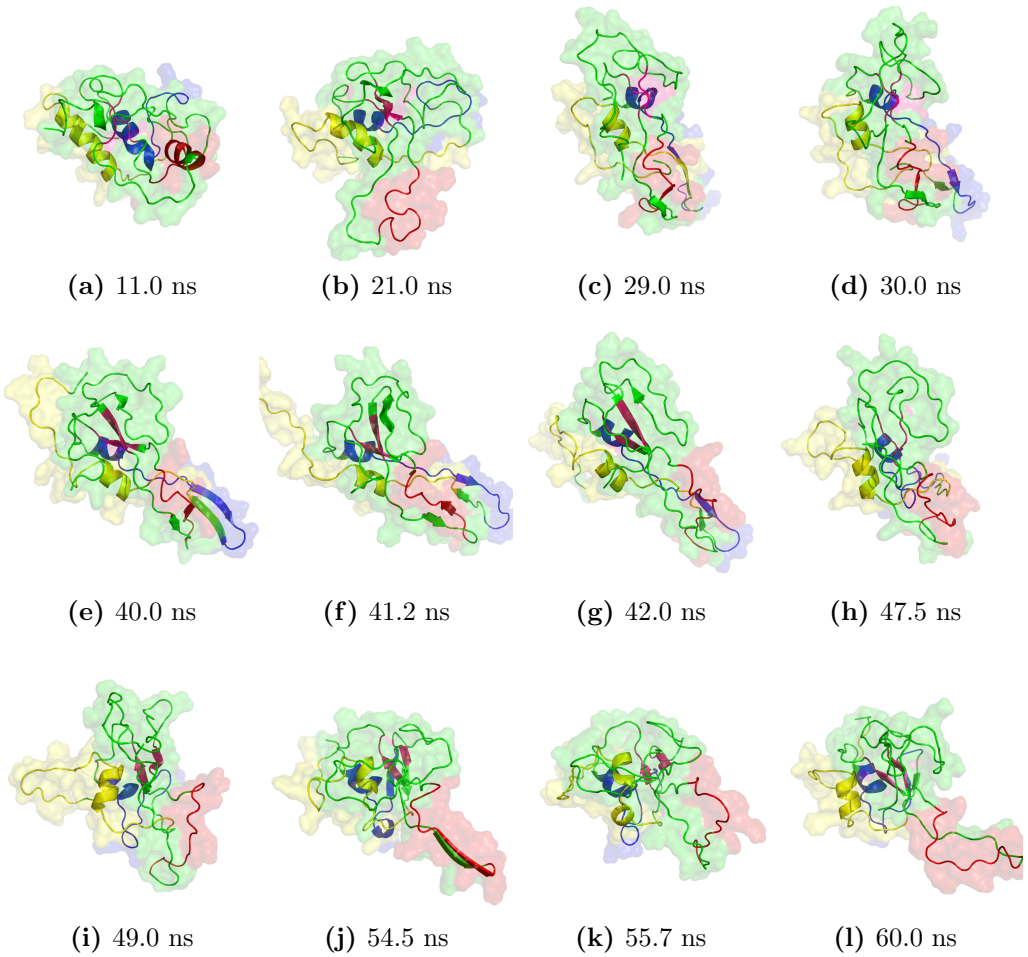
(e) 3.0 ns **(f)** 5.0 ns **(g)** 8.0 ns

Figure 2.10: Starting Snapshots for short simulations. Images adapted with permission of Sara Campos and António Baptista.

60.0 ns; only one replicate per each starting time. All these simulations were performed using the GROMOS 53a6 force field. Additionally, 14 simulations were performed using the GROMOS 43a1 force field, starting from the same points as the long simulations mentioned before.

Constant-pH MD settings

All simulations were performed using the constant-pH MD method implemented for the GROMACS package developed by Baptista et al [**1, 53, 56–62**]. In simulations at pH 7, all 6 His residues were titrated and in the simulation at pH 2 all carboxylic acids (Glu, Asp and C-ter) were titrated. Each constant-pH MD cycle was 2 ps long and the solvent relaxation step was 0.2 ps long.



(a) 11.0 ns **(b)** 21.0 ns **(c)** 29.0 ns **(d)** 30.0 ns

(e) 40.0 ns **(f)** 41.2 ns **(g)** 42.0 ns **(h)** 47.5 ns

(i) 49.0 ns **(j)** 54.5 ns **(k)** 55.7 ns **(l)** 60.0 ns

Figure 2.11: Starting Snapshots for long simulations. Images adapted with permission of Sara Campos and António Baptista.

MM/MD settings

The MM/MD simulations were performed using a modified version of GROMACS 3.2.1 [**57**]. The structures were surrounded by 16089 water molecules (single point charge water molecules [**89**]) in a rhombic dodecahedron box with periodic boundary conditions. The non-bonded interactions were treated using a twin-range cutoff of 8/14 Å and updating the neighbor lists every 10 fs. Electrostatic long range interactions were treated with Generalized Reaction Field [**82**] with a relative dielectric constant of 54 and an ionic

strength of 0.15 M. The Berendsen coupling [**83**] was used to treat temperature (310 K) and pressure (1 bar) with coupling constants of 0.1 and 0.5 respectively. Isothermal compressibility of $4.5 \times 10^{-5}\text{bar}^{-1}$ was used. All bonds were constrained using the LINCS algorithm [**84**].

The minimization procedure had three steps: about 30 steps of Steepest Descent without constraints, 10000 steps of l-BFGS also without constraints and about 10 steps of Steepest Descent with all bonds constrained. The initiation was performed in 5 steps of 50 ps each with different restraints:

1. 1000 kJ.nm^{-1}.mol^{-1} in all protein atoms
2. 1000 kJ.nm^{-1}.mol^{-1} in C_α and 100 kJ.nm^{-1}.mol^{-1} in all other protein atoms
3. 1000 kJ.nm^{-1}.mol^{-1} in C_α and 10 kJ.nm^{-1}.mol^{-1} in all other protein atoms
4. 100 kJ.nm^{-1}.mol^{-1} in C_α and 10 kJ.nm^{-1}.mol^{-1} in all other protein atoms
5. 10 kJ.nm^{-1}.mol^{-1} in C_α

PB/MC settings

The MEAD 2.2.0 [**90**] software package was used for PB calculations. For the atomic charges and radii, the values of GROMOS force field were used. A dielectric constant of 2 for the protein and 80 for the solvent were used. For the finite difference procedure grid spacings of 0.25, 1.0 and 2.0 Å were used. The temperature was 310 K and the ionic strength was 0.15 M.

The MC calculations were performed using the PETIT (version 1.5) [**88**] software with 10^5 steps for each calculation.

2.6 Analysis

Several tools were used from the GROMACS software package [**91**–**93**] and other were developed in-house.

Usually, measurements taken along the simulation time contain large noise. To facilitate the interpretation of the results, one can use a sliding window average. This smoothing technique consists in calculating a series of local average values along the simulation time. These average values are used in the graphic representations instead of the original ones.

It is very useful to look at the conformations obtained from MD and visually check the general features of the system during the simulations. The PyMOL software [**18**] was used in this visualization and, also, to obtain rendered conformation images.

2.6.1 Secondary Structure - DSSP

In 1983, Kabsch and Sander, established a criterion to assign the secondary structure of a protein: the DSSP criterion [**94**]. This criterion is the most used nowadays in protein visualization tools, protein crystallography and NMR. According to DSSP, there are several specific angles and distances characterizing the secondary structure elements. One of the limitations of this method is that a small change in the protein conformation can lead to a large change in secondary structure elements. Nevertheless, there are no other satisfactory criteria to assign the secondary structure. Therefore, in this work, we decided to use the DSSP criterion to evaluate the secondary structure along the simulations.

DSSP calculations were performed using do_dssp tool from GROMACS software package [**91**–**93**] and the dssp program [**94**].

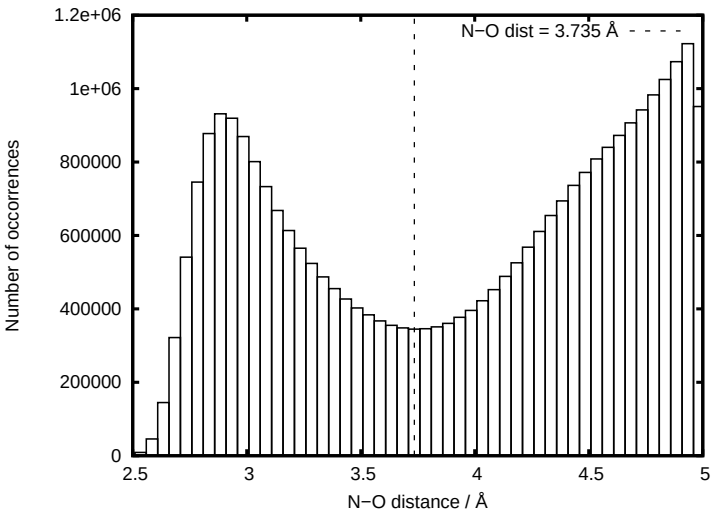


Figure 2.12: Distribution of all N-O distances lower than 5.0 Å in all long simulations performed

2.6.2 Secondary Structure - N-O Contacts Tool

In this work, we developed a tool to analyze the contacts present in a protein over the simulation time. The contact was defined by one pair of N and O atoms from the main chain that are closer than 3.735 Å. This distance was chosen based on the minimum value between the two peaks in the histograms of all N-O distances obtained in the long simulations performed (**Figure 2.12**).

This tool has the advantage of not using a strict criterion like DSSP (section 2.6.1). It is possible to look at the overall of the contacts present in the protein and evaluate the number of contacts kept over the simulation. It is also possible to look in detail at one particular contact that can be of particular importance in a major conformational transition.

All the scripts created for this tool were written with *perl programming language*.

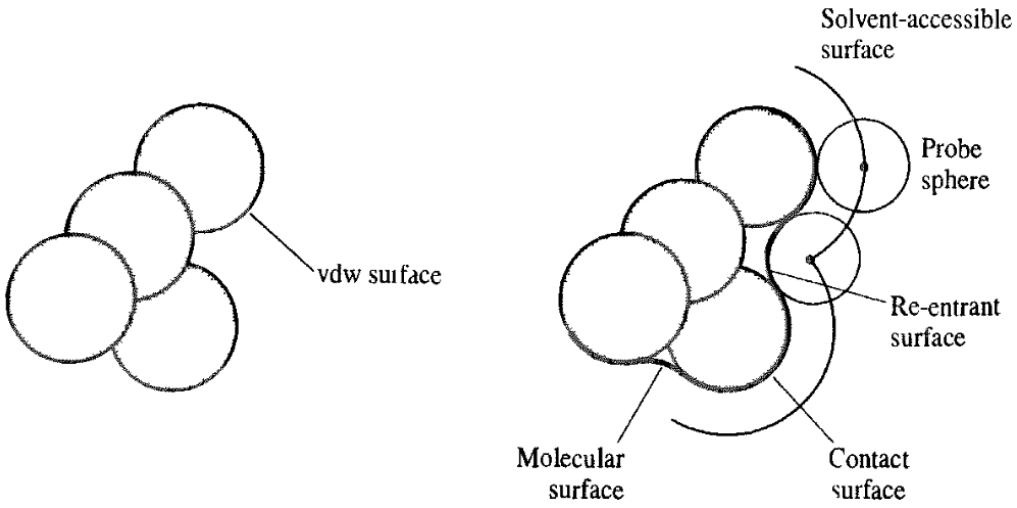


Figure 2.13: Scheme of the algorithm for calculating SAS (adapted from [**54**]).

2.6.3 Solvent Accessible Surface (SAS)

SAS is a measure of how much of the protein surface is in contact with solvent. It is calculated rolling a probe sphere along the van der Walls surface of the protein (**Figure 2.13**). The SAS value is the area that is traced by the center of the rolling probe sphere. In a general way, a globular structure has a lower SAS than an extended one. This property can be useful to follow large conformation transitions that are usually associated with higher exposure to solvent.

SAS calculations were performed using g_sas tool from GROMACS software package [**91–93**].

2.6.4 Radius of Gyration

To have an idea on the compactness of a protein structure, the radius of gyration is often used (**Equation 2.27**).

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} r_i^2 m_i}{\sum_{i=1}^{N} m_i}} \tag{2.27}$$

N is the number of atoms, r_i is the distance between atom i and the center of mass of the molecule and m_i is the mass of particle i. Like SAS, this property can also be used to follow large conformation changes. The radius of gyration and SAS usually show a similar trend which can be useful in protein folding studies.

Radius of Gyration calculations were performed using g_gyrate tool from GROMACS software package [**91–93**].

2.6.5 Root Mean Square Deviation (RMSD)

The most commonly used measurement to evaluate the similarity between two structures is RMSD. This RMSD value can be calculated using **Equation 2.28**. Before the calculation, the two structures are fitted one to another by rotation and translation until the RMSD value is minimum. Usually, one reference structure is selected and the whole simulation trajectory is fitted to it, followed by the RMSD calculation for each frame. The choice of the reference structure is not trivial because a «bad» structure can lead to inconsistent results [**59**]. The choice depends on the purpose of the RMSD calculation. Another decision has to be made regarding which atoms should be used in the RMSD calculation. The RMSD can be calculated for all atoms of a protein, for only the main chain atom, etc.

The non-mass weighted RMSD can be calculated with **Equation 2.28**.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{2.28}$$

N is the total number of atoms being compared and d_i is the distance between the atom i and the same atom in the reference structure (after

fitting).

RMSD calculations were performed using g_rms tool from GROMACS software package [**91–93**].

2.6.6 Principal Component Analysis

Conformational analysis is essential to evaluate if a conformation or a set of conformations is representative or not. It can also help to explain why one specific conformation is more available than another. The main difficulty present in conformational analysis methods, when applied to proteins, is the high flexibility of proteins, rendering it impossible to explore its whole conformational space. Recently, in the Molecular Simulation group at ITQB, was developed a new method that identifies distinct conformation classes based on Principal Component Analysis (PCA) [**95**] applied to conformations obtained from MD simulations [**59**].

It has been shown that best structure for the fitting used in PCA is the central structure [**59**], which is the structure that minimizes the dispersion measure D_i according with **Equation 2.29**.

$$D_i^2 = \frac{1}{n-1} \sum_{j=i}^{n} \mathrm{rmsd}_{ij}^2 \tag{2.29}$$

This RMSD is calculated with the coordinates of all C_α.

Once the central structure is found the PCA step is performed. Given the C_α coordinates of a protein molecule, the number of dimensions is 3N (with N being the number of residues). PCA tries to find a small number of dimensions maintaining the similarity relationships. The objective is to find new vectors ordered by variance retaining most variation in the first ones. These new variables are called Principal Components (PC's). The first PC is the one with highest variation and so forth. This allows to exclude the last PC's and reduce the number of dimensions which describes our system.

After calculating the new coordinate system, it is possible to project our data in a two or three dimensional space and obtain a visual representation of the clusters present in our data set. With this representation it is possible to estimate the density in each point of the space. For this, it can be used a Gaussian kernel estimator [**96**]. The kernel bandwidth (h) is defined by **Equation 2.30**.

$$h = \sigma \left(\frac{4}{n(2d+1)} \right)^{1/(d+4)} \tag{2.30}$$

σ^2 is the average marginal variance, n is the number of data points and d is the number of dimensions. After performing this procedure, a probability density function ($P(\mathbf{r})$) is defined of each point in the d-dimensional space.

An energy landscape can be calculated from $P(\mathbf{r})$ with **Equation 2.31**.

$$E(\mathbf{r}) = -RTln\frac{P(\mathbf{r})}{P_{max}} \tag{2.31}$$

P_{max} is the maximum of $P(\mathbf{r})$. Several examples of energy landscapes are shown in section 3.5. The main limitation of these landscape is that two zones that are separated in the multidimensional space can be overlapped in the 2D space.

Chapter 3

Results and Discussion

In the present chapter will be described and discussed the most relevant results of our simulations. Before that, it is necessary to clarify the used nomenclature:

- Long simulations - equilibrated segment of long simulations (20 - 40 ns) performed for this thesis whose setup were already described in section 2.5
- Short simulations - 10 ns simulations performed for this thesis whose setup were already described in section 2.5
- pH 2 simulations - set of 12 simulations previously performed at pH 2 in Molecular Simulation Group at ITQB [**1**]
- pH 7 simulations - set of 12 simulations previously performed at pH 7 in Molecular Simulation Group at ITQB [**1**]
- pH2mis - simulation from which our simulations were started (pH 2). This simulation is one of the pH 2 simulations set and it was already discussed in sections 1.2.2 and 2.5

- pH7mis - simulation performed with the same initial structure and initiation and minimization procedures as pH2mis but at pH 7. This simulation is one of the pH 7 simulations set.

3.1 Secondary Structure - DSSP

The secondary structure was computed for each sampled conformation of all simulations performed. **Figure 3.1** represents the variation of the total number of residues in helix for short simulations. In this figure, it is also represented the variation of helicity for pH2mis and pH7mis simulations. Through the comparison of the helix content of short simulations with the reference curves (pH2mis and pH7mis) it is possible to evaluate the initial trend of the helix content after the change in pH to 7: if its going down like pH2mis simulation or if it tends to stabilize like in pH7mis simulation. There are replicates with the three possible behaviors: the helix content increases, decreases or stabilizes. Nevertheless, **Figure 3.1** shows that the global trend of the helix content is to decrease even after the change to pH 7. However, the helix content of simulations started at $t = 0.5,\ 1.0,\ 1.5$ and 2.0 ns is higher than in the pH2mis one.

In **Figure 3.2** it is shown a secondary structure map of a long simulation (starting time: 30.0 ns). The helix content is low and unstable. Moreover, HA (aa 144-156) had been lost irreversibly in pH2mis simulation. This observation is valid for all long simulations (data not shown) except for the one that started from $t = 11.0$ ns (HA was not yet lost in pH2mis simulation). In the particular replicate represented in **Figure 3.2**, HB (aa 174-194) becomes unstable at the end of the simulation. This is a common observation in other replicates for both HB and HC: in some simulations they were unstable and, occasionally, completely lost. On the other hand, there is a high β content in all long simulations and the β structures originally present in the starting conformations are very difficult to disrupt. Moreover, there

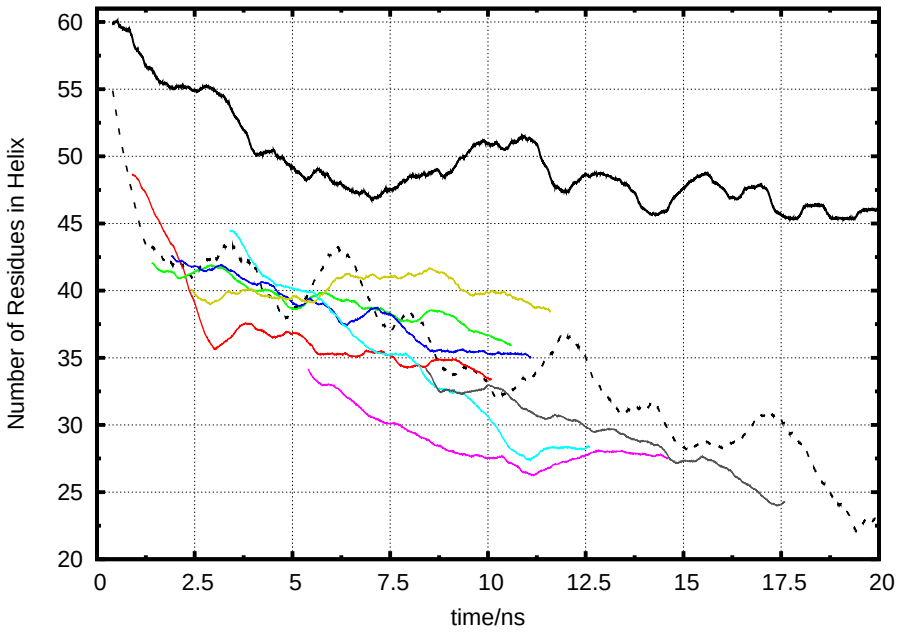


Figure 3.1: Variation of Helix content in short simulations. — pH7mis - - pH2mis. The remaining lines correspond to short simulations. For these, the initial time is shifted to the pH2mis time of the conformation from which they started. It is shown an average value of all three replicates per each starting time. — 0.5 ns — 1.0 ns — 1.5 ns — 2.0 ns — 3.0 ns — 5.0 ns — 8.0 ns.

are several new β structures that are meta-stable [1] appearing in different regions of the protein.

Figure 3.3 shows the ratio of helix content in pH2mis, pH7mis and the average values for long simulations per residue. The helix content is much lower in the long simulations than in the pH7mis one. This observation is in agreement with the results from **Figure 3.2**. However, the comparison

[1] Here, meta-stable means that the referred structural feature is formed and disrupted several times along the simulation and in many regions of the protein.

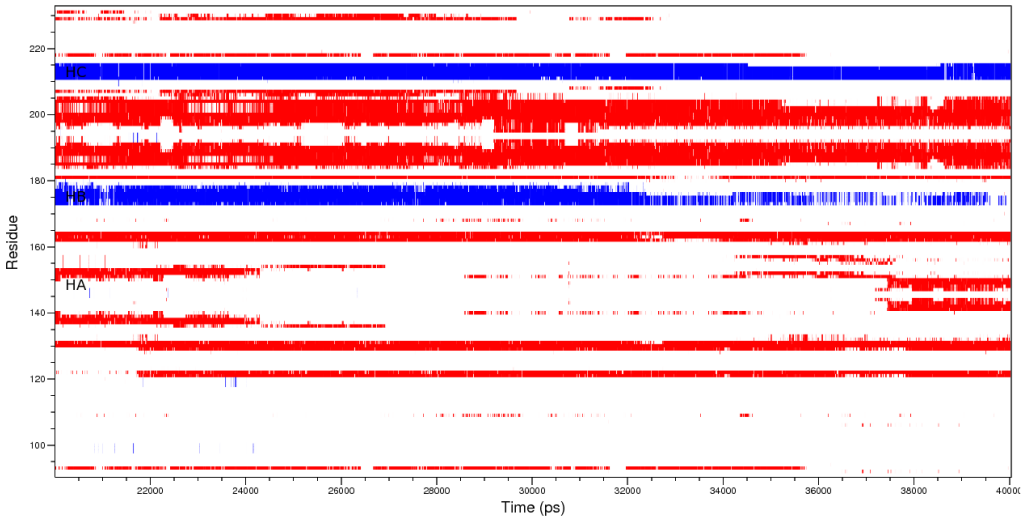


Figure 3.2: Secondary Structure map of a long simulation (starting time: 30.0 ns). ■ represents residues in β-structure, ■ represents the residues in helices.

between pH2mis and long simulations leads to different observations. In the region of HA, the long simulations have a higher ratio of helix content than pH2mis. In the regions of HB and HC the opposite happens.

The secondary structure analysis of our simulations do not show any kind of reversibility in the misfolding transition of PrP. However, there are several factors that can contribute to the considerable lowering in the helix content of our simulations which could mislead us into wrongfull interpretations. The abrupt change in the pH, and the consequent addition of several new negative charges, can lead to new unfavorable interactions which can disrupt the global shape of a helix. To minimize this effect, we used a smooth minimization / initiation protocol in our simulations, but it is difficult to completely avoid these instabilities. Nevertheless, even small effects can have bigger consequences since the DSSP criterion is strict and sensitive to small changes in distances or angles. Another factor is that the time it takes to form new helix structures is unknown and it can take longer than a μs which renders the calculations prohibitive. The high content in β structures

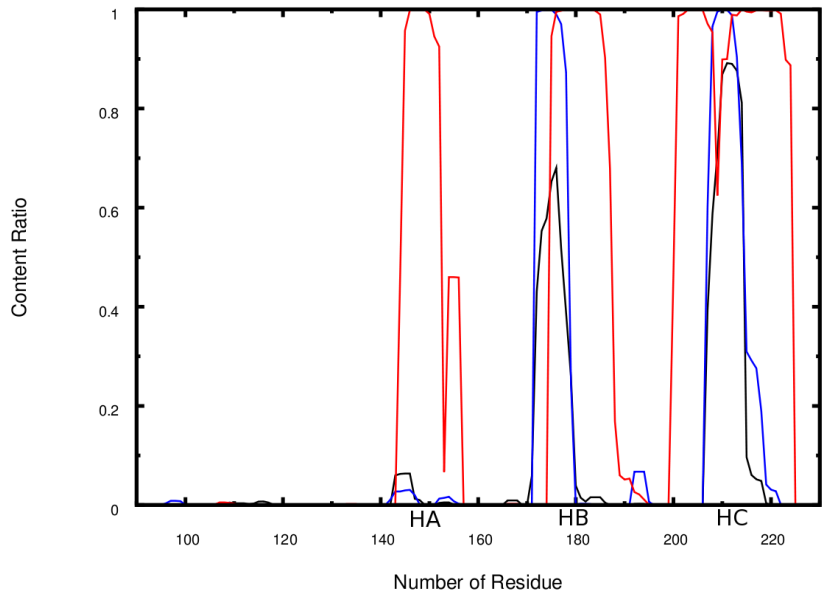


Figure 3.3: Ratio of helix content in pH2mis, pH7mis and long simulations per residue. – pH2mis, – pH7mis, – Average values of long simulations.

of our long simulations seems to suggest that this structures are also stable at pH7. In this kind of long simulations, we should also take into account the fact that force fields can destabilise the secondary structure of peptides [**79**] and proteins [**78**].

All these factors, together with the unnatural loss of helix content in our simulations at pH 7, led us to search for alternative ways of measuring our protein fold.

3.2 Secondary Structure - N-O Contacts

We developed a N-O Contacts tool (explained in section 2.6.2) in order to overcome some of the limitations associated with the strict DSSP criterion. We think that this measure can attenuate the drastic effects observed in the secondary structure of our protein because it does not take into account

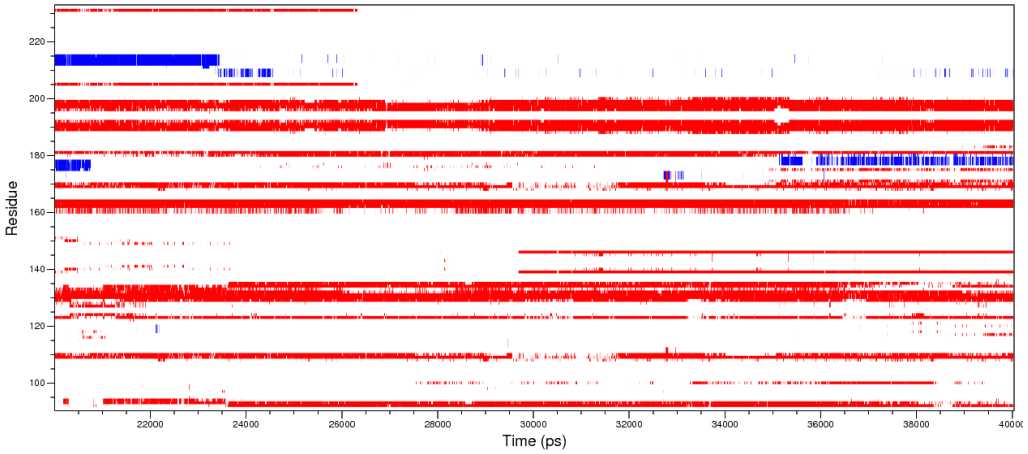
the angle of interaction. In order to compare the results of this tool with the ones obtained with the DSSP criterion we established the difference between short and long range contacts:

- Short range - 1:3, 1:4 or 1:5 contact (comparable to helix structures) - Contact involving atoms which are 2, 3 or 4 residues apart
- Long range - contact involving atoms which are 5 or more residues apart (comparable to β structures)

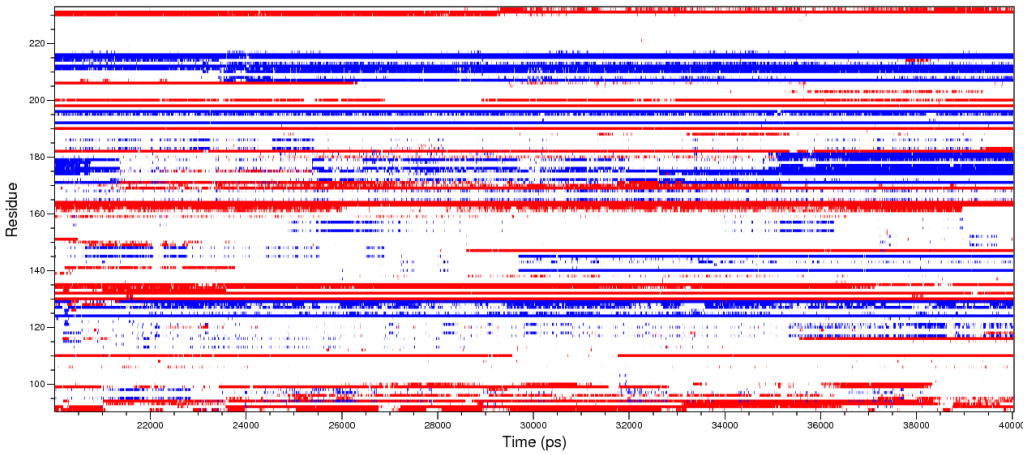
In **Figure 3.4** the secondary structure map (obtained from DSSP) and the N-O contacts variation for a long simulation (29.0 ns) are represented. In this particular replicate the helix content is completely lost at $t \simeq 25.0$ns and partially recovered at $t \simeq 35.0$ns. As expected, this drastic effect is attenuated in the analysis with the N-O contacts tool. The content in long range contacts is approximately the same as in β structures. However, the content in short range contacts is higher than in helix structures obtained from DSSP. Despite the instability observed in the helices, the content in short range contacts is always present. Therefore, these results suggest that the global shape of the helix is maintained despite the observed lowering in DSSP's helix content. **Figure 3.5** shows a representation of two short range contacts which are not considered as a part of a helix. In the cartoon representation it is evident that the global shape of a helix is preserved, although this contact is not considered as a helix according to the DSSP criterion. These results show that the conformational changes are smaller than what we previously induced from the DSSP results.

This tool also allowed us to evaluate the reversibility from a different point of view: assessing important contacts. We define that an important contact is present in a reference simulation when its occurrence is over 10%. Once defined the important contacts in pH2mis (**imp-ph2**) and pH7mis (**imp-ph7**) simulations, we determined the number of these contacts present in short simulations [2] (**Figure 3.6** and **appendix A**). This type of analysis

[2]The initial 10 ns of the long simulations can be treated as short simulations.

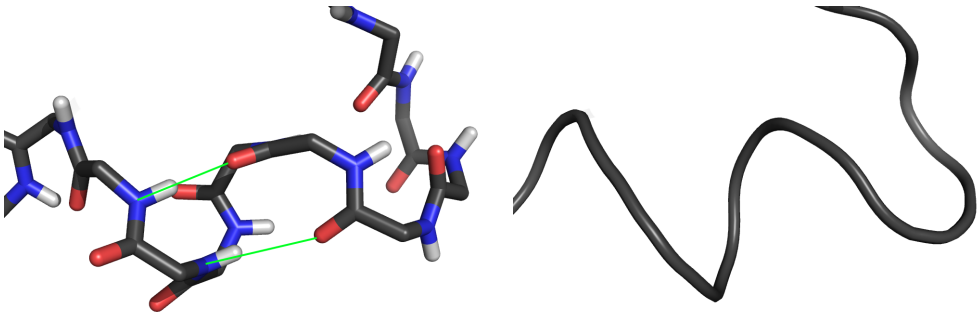


(a) DSSP map



(b) N-O contacts variation

Figure 3.4: Comparison between variation of secondary structure according to the DSSP criterion and N-O contacts tool in a long simulation (29.0 ns) ■ β structure or long range contact, ■ helix structure or short range contact.

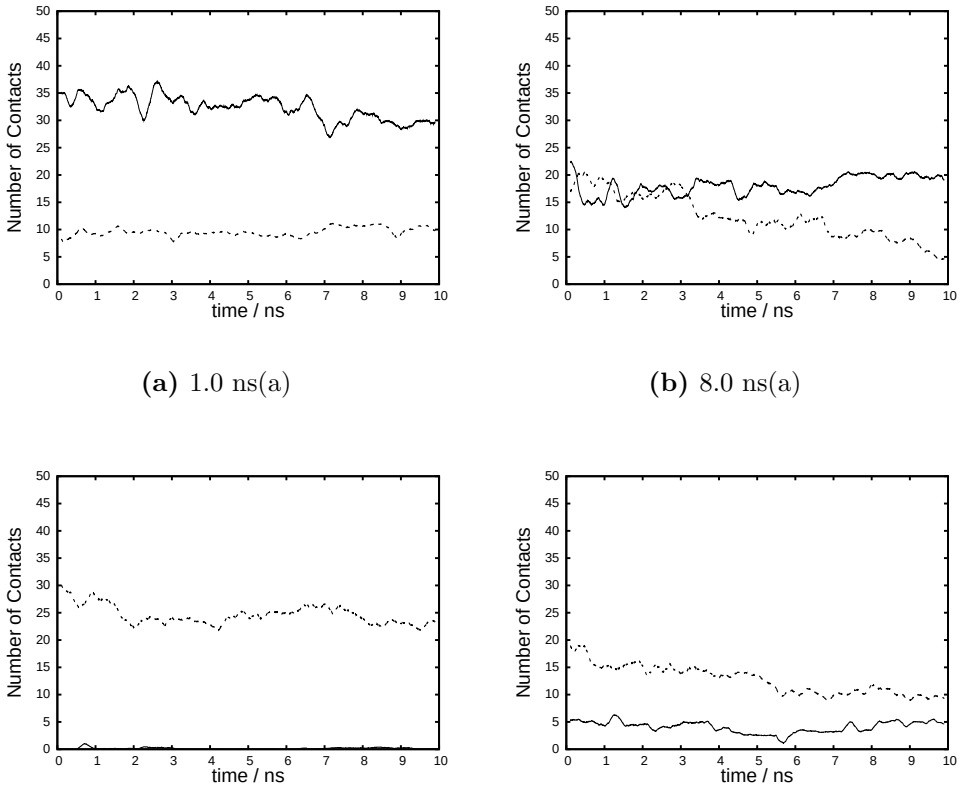


(a) Main Chain atoms **(b)** Cartoon representation

Figure 3.5: Structural detail of two short range contacts (–) which are not considered as a part of a helix.

allows us to evaluate the initial trend after the change in pH to 7. In the first three simulations (**Figure A.1a**, **A.1b** and **A.2a**) there is a decrease in imp-ph7 contacts along the simulation time. This observation may be due to the fact that the change in pH to 7 was made at an early time in the pH2mis simulation. At this time the protein was not yet adapted to the force field. In this case, there were two factors contributing to the destabilization of the protein: the adaptation to the force field [3] and the abrupt change in the pH. As expected, the simulations started from earlier times have more imp-ph7 contacts than imp-pH2 ones. Between 8.0 ns and 11.0 ns this trend is reversed. In the simulations started at $t > 21.0$ ns the number of imp-ph7 contacts that were maintained is almost null, meaning that these conformations are very altered and it is very difficult to recover a significant number of imp-ph7 contacts. Moreover, in a general way, there are no evident transitions after the change to pH 7 and the curves are approximately stable in most of the simulations. However, in a simulation started at $t = 8.0$ ns (**Figure 3.6b** and **A.4e**) there was a transition: in the beginning the number of imp-ph2 and imp-ph7 contacts is approximately the same. Along the simulation time the number of imp-ph7 contacts increase while the imp-ph2 ones decrease. This simulation indicates some

[3] Despite the use of an initiation procedure, it is necessary to wait some simulation time (usually ~ 1 ns) to let the protein equilibrate and adapt to the force field.



(a) 1.0 ns(a)

(b) 8.0 ns(a)

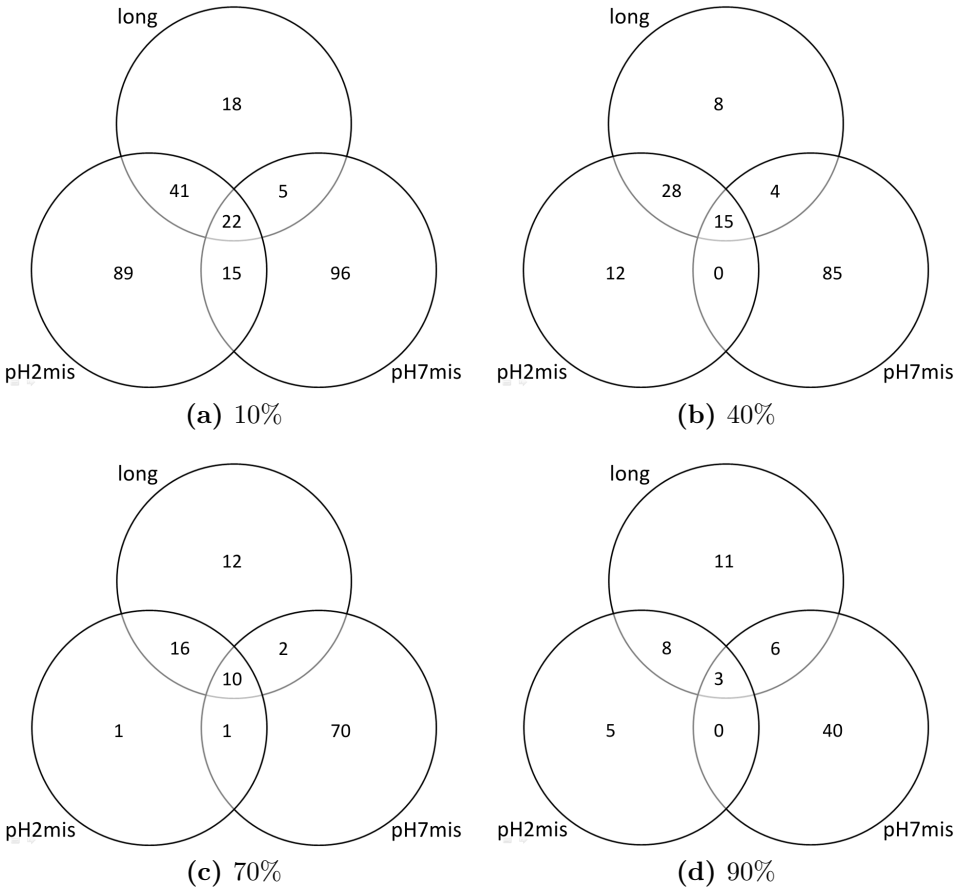
(c) 40.0 ns

(d) 60.0 ns

Figure 3.6: Variation of the number of important contacts maintained in short simulations. imp-pH7 (—), imp-pH2 (··).

reversibility in the early steps of the misfolding transition of PrP.

We also decided to study the recovery of important contacts lost in the misfolding transition. We defined another set of important contacts for our long simulations (**imp-long**) as a contact present with an occurrence over 10% in at least two long simulations. With the three defined sets (imp-ph2, imp-ph7 and imp-long) it is possible to plot them in a Venn Diagram (**Figure 3.7a**). We also tested different cut-off values for contact occurrence (**Figure 3.7**).



(a) 10% **(b)** 40% **(c)** 70% **(d)** 90%

Figure 3.7: Venn Diagrams for the three defined sets of contacts: imp-ph2, imp-ph7 and imp-long. Cutoff values for contact occurrence are indicated.

When using a cutoff of 90 %, we identify the most important contacts available. In this case, there is a large number of contacts specific of pH7mis simulation (**Figure 3.7d**). These contacts are lost irreversibly in pH2mis simulation. A small number of contacts are present in all sets. In **Figure 3.8** are represented the three contacts present in all sets with an occurrence over 90%. Two of these contacts are in the region of the β-sheet originally present in PrP and the third one is in the HB close to the disulfide bond. This result ilustrates the two most stable regions of PrP at different pH values. Only at a low cutoff (10%) there is a large number of contacts ex-

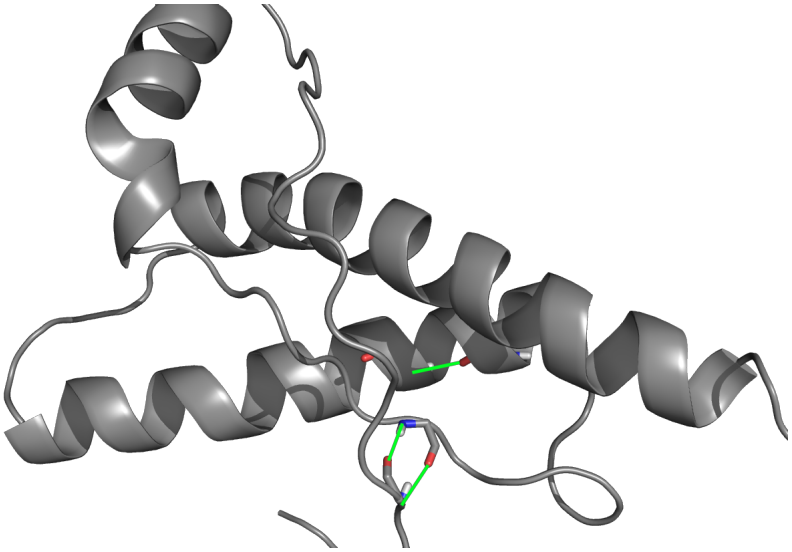


Figure 3.8: Structural detail of the three contacts present in pH2mis, pH7mis and long simulations (—) with an occurrence over 90%. Starting snapshot of pH2mis and pH7mis simulations is used.

clusive of the pH2mis simulation. This is in agreement with the observation that the new formed structures are meta-stable. The contacts present in both pH2mis and our long simulations are the ones formed at pH 2 and that were not disrupted when pH was changed to 7 (this group of contacts has a large number of elements for the 10% cutoff which supports the previous observation of meta-stability). Finally, there is a small number of contacts with an occurrence over 90% that are present in both the pH7mis and our long simulations (**Figure 3.9**) (in addition to the 3 contacts shown in **Figure 3.8**). These six contacts are located in HB and HC, and can be seen as the recovered contacts when pH was changed to 7.

The results obtained with this tool are different from the ones obtained with the DSSP. In one hand, the N-O contacts tool shows the meta-stability of long range contacts (such as DSSP) but on the other hand, it also suggests that the short range contacts are stabilized by pH 7. In some cases, these short range contacts are not helices according to the DSSP criterion, however, they are captured with this tool. Therefore, we think that the N-O contacts tool is more robust than DSSP and less sensitive to small

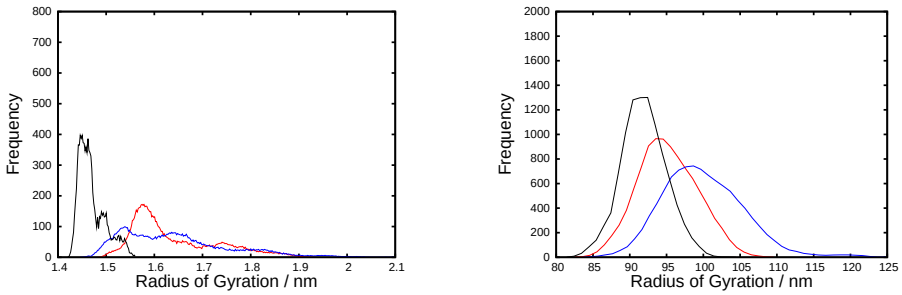
Figure 3.9: Structural detail of the zones involved in the six contacts present in pH7mis and long simulations (■) with an occurrence over 90%. Starting snapshot of pH2mis and pH7mis simulations is used.

conformational changes.

The results obtained with the present tool do not show clear reversibility events. However, in some simulations, they reveal the correct trend needed for the pH induced reversibility to occur. Nevertheless, even though secondary structure is generally a good indicator of the fold state of a protein, we think that there are other structural measures that can be very helpful in determining the overall fold state of PrP. In the next sections, we will explore four different methodologies: radius of gyration, SAS, RMSD and PCA.

3.3 Radius of Gyration and SAS

The radius of gyration and solvent accessible surface (SAS) are two measures of the global shape of a protein. However, the interpretation of the temporal variation of these properties is very difficult mainly because their



(a) Radius of Gyration **(b)** SAS

Figure 3.10: Histograms of Radius of Gyration and SAS of a long simulation. – Average values of pH 2 simulations, – Average values of pH 7 simulations, – Simulation started at $t = 11.0$ ns.

measured value is the result of a large number of factors. To simplify the interpretation, we decided to look at the histograms of these properties and compare them with the reference ones (average values for pH 2 and pH 7 simulations) (see Appendix B). Also, and because these are equilibrium properties, we only studied the long simulations.

The histogram of the long simulation started at $t = 11.0$ ns is represented in **Figure 3.10**. The values of the radius of gyration and SAS of the pH 2 simulations present more variety than the values for the pH 7 simulations. This confirms our previous observations that the pH 2 simulations have more conformational variability than the pH 7 ones. The long simulation (**Figure 3.10**) presents values lower than both reference ones. This means that the protein is in a more compact conformation for this simulation. The results in Appendix B show that the distributions strongly depend on the starting conformation. In the simulations started in the early times, the values are small but with a tendency to increase. In the simulations started at 30.0, 40.0, 41.2 and 42.0 these values are higher than the reference ones. This means that the protein acquired less compact conformations. The following simulations (47.5 and 49.0 ns) present values in the same range of the reference ones. Finally, the simulations started at 54.5, 55.7 and 60.0

ns present values of radius of gyration lower than the reference ones but the SAS values are in the same range of the reference ones. One possible interpretation for these is that the protein acquired a more compact fold (closer to a globular shape) but kept several internal regions in contact with the solvent. These results confirm the general notion that the long simulations at pH 7 are very dependent on the starting conformation and the change in pH induces some type of stabilization (see range in the distributions present in appendix B). Also, for simulations that start from conformations taken at times as early as 21 ns, the protein has already experienced such drastic conformational change that it would be very improbable to observe any type of reversibility in our timescale.

We also plotted the radius of gyration vs SAS and these two properties vs RMSD (data not shown) in order to obtain energy landscapes similar to the ones present in section 3.5. However, the obtained landscapes did not allow a good separation between the three groups of simulations (pH 2, pH 7 and our long simulations).

3.4 RMSD

RMSD is the measure that can, in principle, overcome some of the limitations we observed with radius of gyration and SAS.

As explained in section 2.6.5, the RMSD value is calculated against a reference structure. For this analysis, we chose two references, namely the central structures (see section 2.6.6) from the simulations at pH 2 and pH 7. The RMSD calculations were also done for all short simulations to evaluate the initial trend after the change in pH. The RMSD values of our simulations against the central structure of pH 2 and pH 7 were also plotted against each other in order to evaluate if the simulations were closer to pH 2 or pH 7. If the value is above (below) the x=y line the conformation is closer to pH 2 (pH 7). Moreover, it is possible to evaluate if, during the first 10 ns, the protein is getting closer to pH 2 or pH 7.

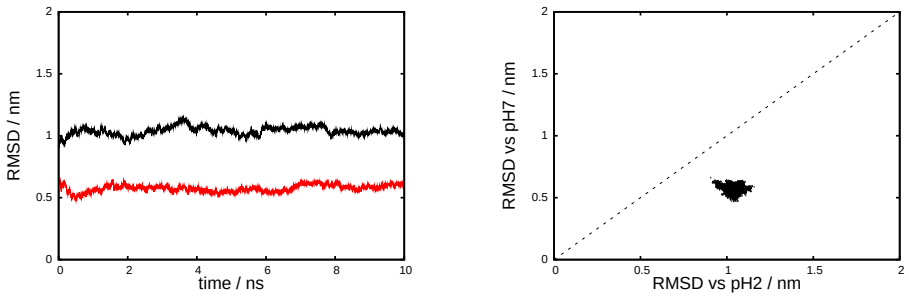


Figure 3.11: Simulation: 0.5 ns(c). See caption in **Figure 3.16**.

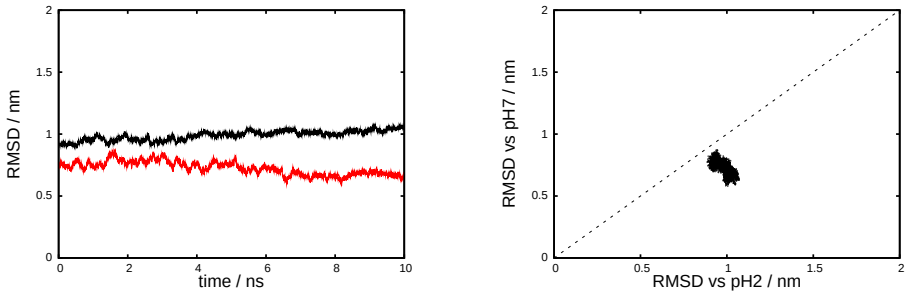


Figure 3.12: Simulation: 1.5 ns(a). See caption in **Figure 3.16**.

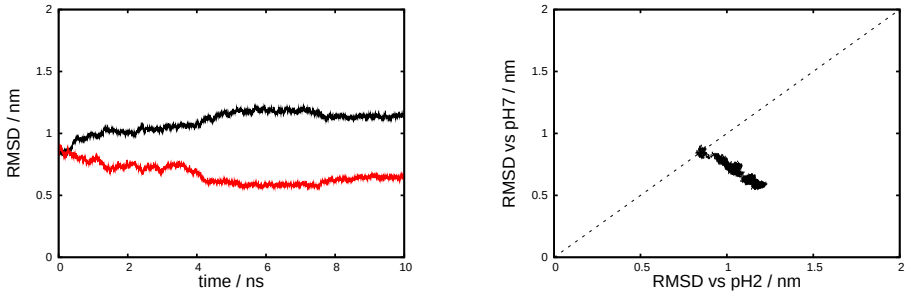


Figure 3.13: Simulation: 3.0 ns(a). See caption in **Figure 3.16**.

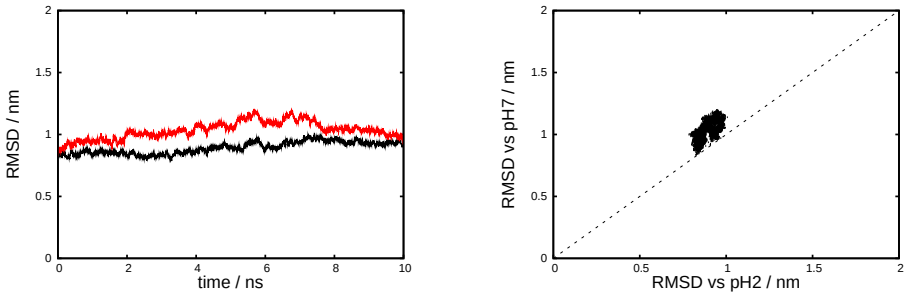


Figure 3.14: Simulation: 3.0 ns(c). See caption in **Figure 3.16**.

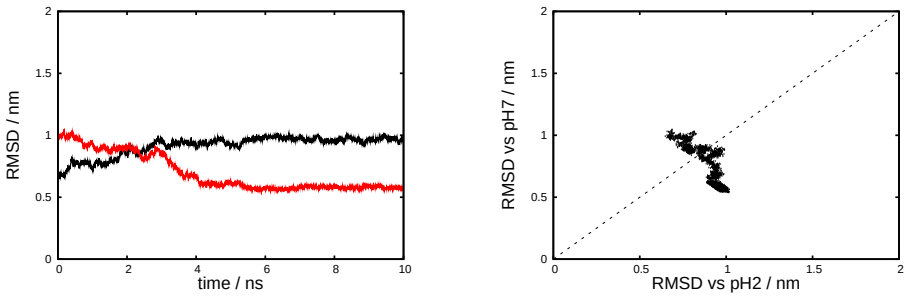


Figure 3.15: Simulation: 8.0 ns(a). See caption in **Figure 3.16**.

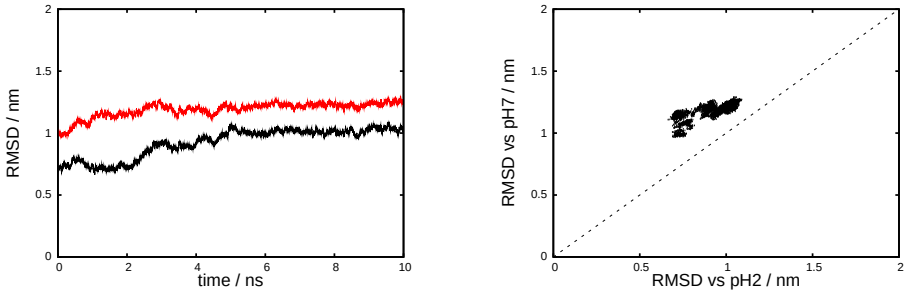


Figure 3.16: Simulation: 8.0 ns(b). left) – RMSD vs average structure of pH 7 simulations, – RMSD vs average structure of pH 2 simulations. **right)**, • (RMSD vs average structure of pH 7 simulations) vs (RMSD vs average structure of pH 2 simulations), ·· x=y line.

Figures 3.11 - 3.16 and Appendix C show the described RMSD plots. In these graphic representations, different behaviors occur after the change in pH to 7 (starting times are written in the beginning of each item):

- **0.5a - 1.5c** (**Figures C.1 - C.9**, **3.11** and **3.12**) - In these simulations, the RMSD values are always below the bisection which means that the protein is closer to pH 7. Also, the temporal curves present low fluctuations and the dots are restricted to a small space. This corroborates that the pH 7 had a stabilizing effect on the protein conformation. In simulations **0.5b**, **1.5a** and **1.5b** the RMSD against central structure of pH 7 slightly decrease along the simulation time and the opposite happens with the RMSD against the pH 2 one. This observation might suggest a partial reversibility event occurring in these simulations.

- **2.0a**, **2.0b**, **3.0a**, **5.0a**, **8.0a**, **8.0c** (**Figures C.10**, **C.11**, **C.13**, **3.13**, **C.16**, **C.19 3.15**, and **C.21**) - In all these simulations the RMSD values also suggest that a partial reversibility event occurred. The RMSD against the central structure of pH 7 decrease along the simulation time and the opposite happens with the RMSD against the pH 2 one.

- **2.0c**, **3.0b**, **3.0c**, **5.0b**, **5.0c** and **8.0b** (**Figures C.12**, **C.14**, **C.15**, **3.14**, **C.17**, **C.18**, **C.20** and **3.16**) - In this group of simulations the RMSD values over time do not present a clear trend. In some cases, the simulation is moving away from both central structures (pH 2 and pH 7). These values are always close to x=y line and restricted to a small space. These observations suggest that the pH 7 also had a stabilizing effect in these simulations.

- **11.0** (**Figure C.22**) - Although the RMSD values of these simulations are always above the bisection, the protein is getting closer to the reference structure of pH 7. The RMSD against central structure of

pH 7 slightly decreases along the simulation time and the opposite happens with the RMSD against the pH 2 one.

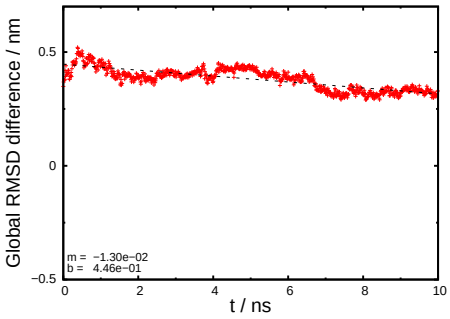
- **21.0 - 60.0** (**Figures C.23 - C.33**) - In a general way, the RMSD values for these short simulations are high, which means that the protein is too far from both reference structures.

Since the presented RMSD calculations were done against only two representative conformations, their interpretation might contain significant uncertainty. In order to minimize this, we decided to calculate the average value of the RMSD against 1000 conformations of each reference simulation (pH2mis and pH7mis). Then, we calculated the difference between these two values (pH2 - pH7) per each conformation in the short simulations. If the calculated difference is positive the protein is closer to pH7mis simulation. Also, it is possible to quantify this trend if we do a least-square fit of these differences over time. If the calculated slope is positive (negative) the protein is becoming closer to the pH7mis (pH2mis) simulation.

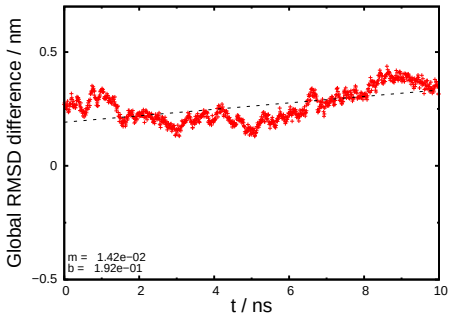
The graphic representations obtained with the described methodology are shown in **Figure 3.17** and Appendix D. The three possible scenarios are well represented in our simulations:

- The slope of the trendline is positive in 1.5a, 1.5b, 2.0a, 2.0b, 3.0a, 8.0a, 8.0b, 8.0c and 11.0 simulations
- The slope of the trendline is approximately zero in 0.5b, 0.5c, 1.0a, 1.0b, 1.0c, 1.5c, 2.0c, 3.0b, 5.0a, 5.0b and 5.0c simulations
- The slope of the trendline is negative in 0.5a and 3.0c simulations

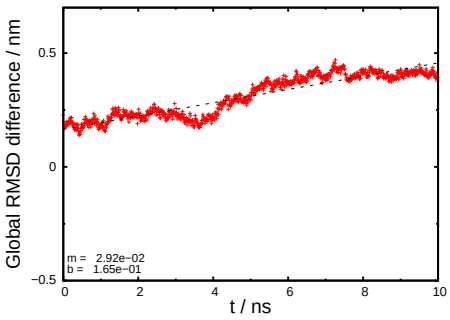
In most of the simulations the slope of the trendline is approximately zero which means that in these simulations there is no clear trend towards any of the typical conformations of pH 2 or pH 7. There is also a significant number of simulations in which the slope of the trendline is positive, showing



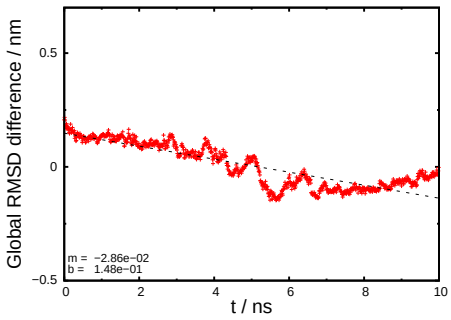
(a) 0.5 ns(c)



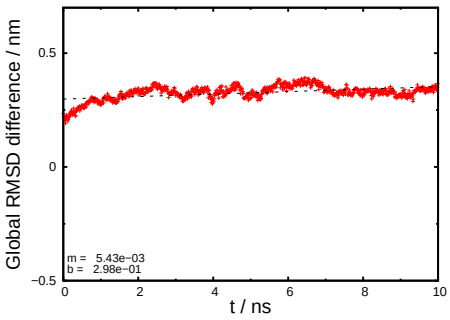
(b) 1.5 ns(a)



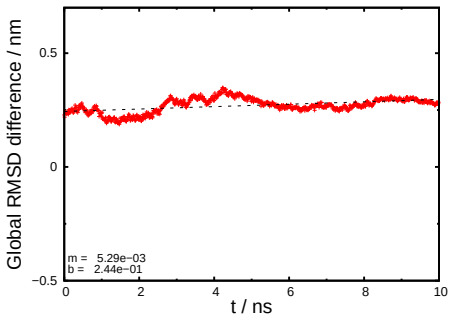
(c) 3.0 ns(a)



(d) 3.0 ns(c)



(e) 8.0 ns(a)



(f) 8.0 ns(b)

Figure 3.17: Difference of average RMSD value vs pH2mis and pH7mis conformations in the short simulations (+) and trendline (- -).

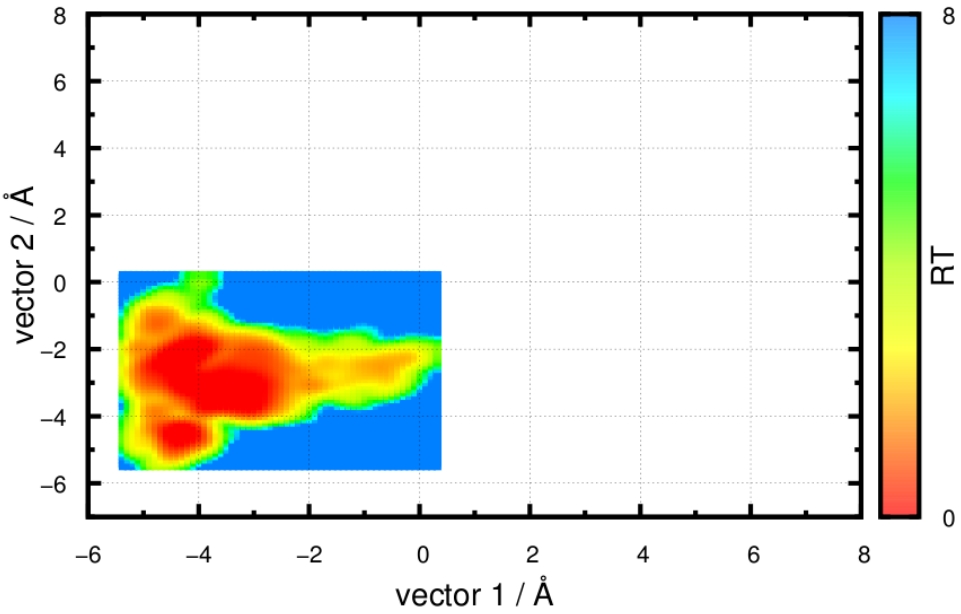
evidence for a possible reversibility event. Interestingly, in simulations 0.5a and 3.0c, the change in pH (to 7) was not able to avoid the transition towards the typical conformations of pH 2.

The RMSD results strengthened our previous hints that reversibility is indeed possible. However, it is more probable in simulations started at early times of the pH2mis simulations. This indicates that the most significant alterations in the misfolding transition might have occurred early in the pH2mis simulation.

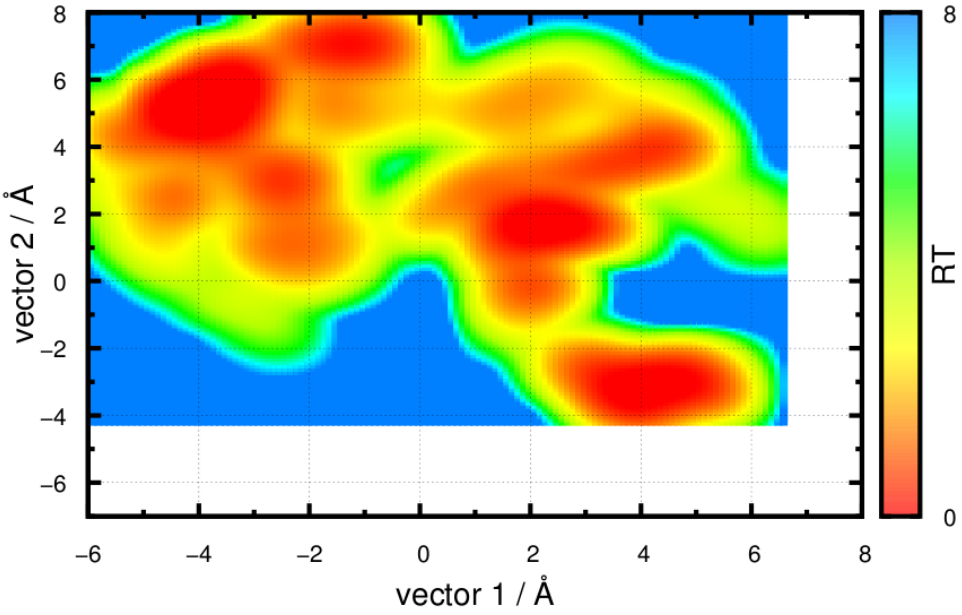
3.5 PCA

Finally, we used Principal Component Analysis (PCA) to clarify possible reversibility events in our simulations, suggested by previous analysis. The PCA calculations were done with the equilibrated parts of 36 simulations (12 long simulations from the present work, 12 at pH 2 and 12 at pH 7 both from previous work in Molecular Simulation Group at ITQB [**1**]). The procedure of this analysis is briefly described in section 2.6.6 and with more detail in ref. [**59**]. In the present section we used the first two Principal Components (PC's) to obtain all the landscapes. To have a 100% correct PCA it is necessary to have a complete sampling of the whole conformational space of the protein and, nowadays, such task is impossible. Nevertheless, a PCA can help to understand this type of processes if it is considered carefully.

With this methodology it is possible to obtain two energy landscapes at pH 7: one using the pH 7 simulations and another one using our long simulations (**Figure 3.18**). One of these energy landscapes (**Figure 3.18a**) has its single basin approximately at $(-4, -3)$. The other one (**Figure 3.18b**) has many basins scattered over the energy landscape. This may indicate that our long simulations have more conformational variability than the previous ones at pH 7. This is not unexpected due to the large variety of starting conformations in our long simulations. Moreover, these two energy

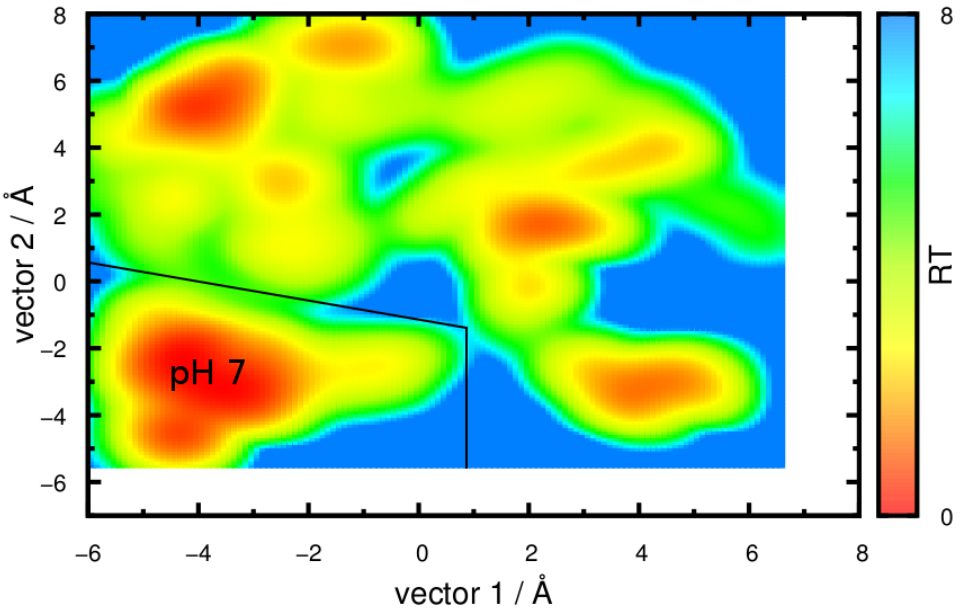


(a) pH 7 simulations

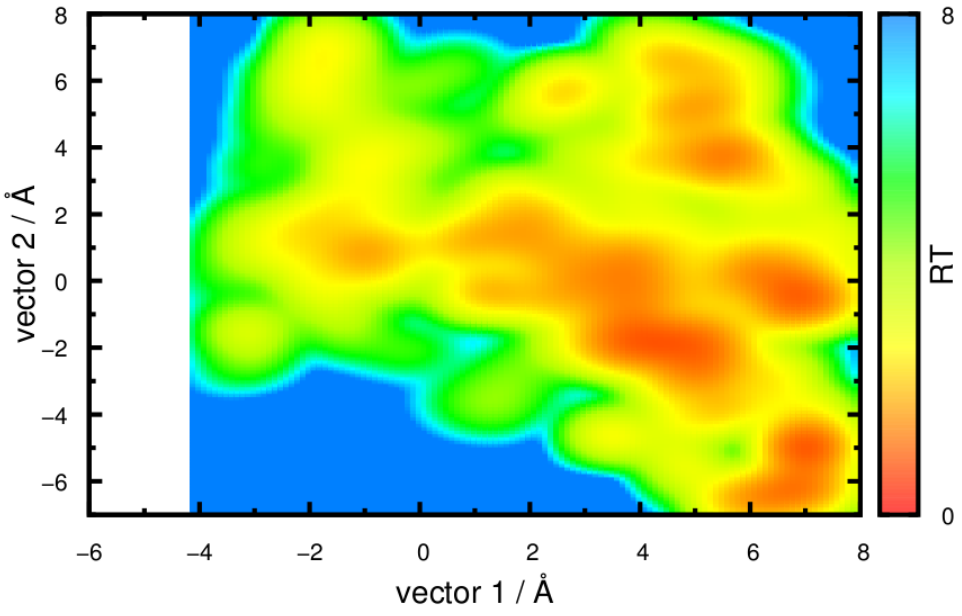


(b) Long simulations from present work

Figure 3.18: Energy landscapes at pH 7 using the first two PC's.



(a) pH 7



(b) pH 2

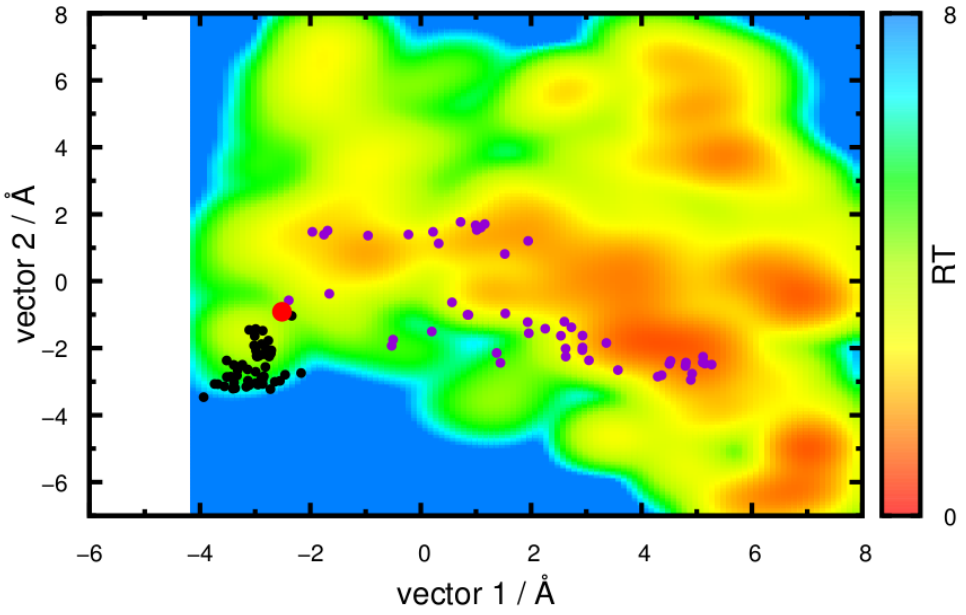
Figure 3.19: Energy landscapes at pH 7 and pH 2 using the first two PC's.

landscapes are completely separated. Therefore, we decided to create a single landscape using both the pH 7 and long simulations (**Figure 3.19a**). In this energy landscape we can not know the relative depth of the observed basins since we did not observe transitions between the two landscapes. However, the obtained landscape should be a decent overall representation at pH 7.

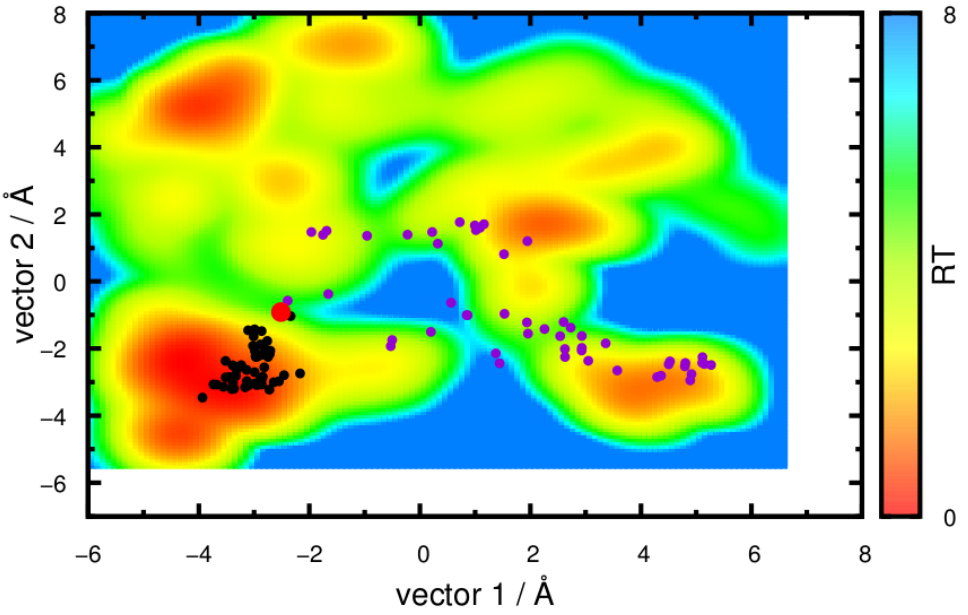
Using the described procedure, it was possible to obtain two energy landscapes: one at pH 7 and another one at pH 2. These landscapes are represented in **Figure 3.19**. The landscape at pH 2 (**Figure 3.19b**) also has many basins scattered over the energy landscape which means that, at this pH value, the protein has a large conformational variability. The basins corresponding to the long simulations are also scattered over the energy landscape. However, the deeper basins change position after the change in the pH to 7. This observation indicates that, after the change in pH, there were new populated regions that were not available at pH 2. Moreover, in these landscapes the normally folded pH 7 region is clearly separated from the other zones, rendering these landscapes a good tool to evaluate reversibility events.

Therefore, we could use these landscapes to evaluate the reversibility in the short simulations. In **Figures 3.20 - 3.25** and Appendix E are represented the same energy landscapes of **Figure 3.19** together with the projection of representative snapshots of the short simulations and the equivalent ones of the pH2mis simulation. For example, in simulations started at $t = 5.0$ ns we used 50 snapshots (50 ps, 100 ps, 150 ps, etc., until 10.0 ns) and the corresponding snapshots in the pH2mis simulation (5050 ps, 5100 ps, 5150 ps, etc., until 15.0 ns). In a general overview, the simulation at pH 2 tends to spread over the landscape more than the short simulations after the change in pH to 7. This observation reinforces the idea that pH 7 has a stabilizing effect.

Next, we briefly describe the projections of each short simulation in the landscapes (starting times are written at the beginning of each item):

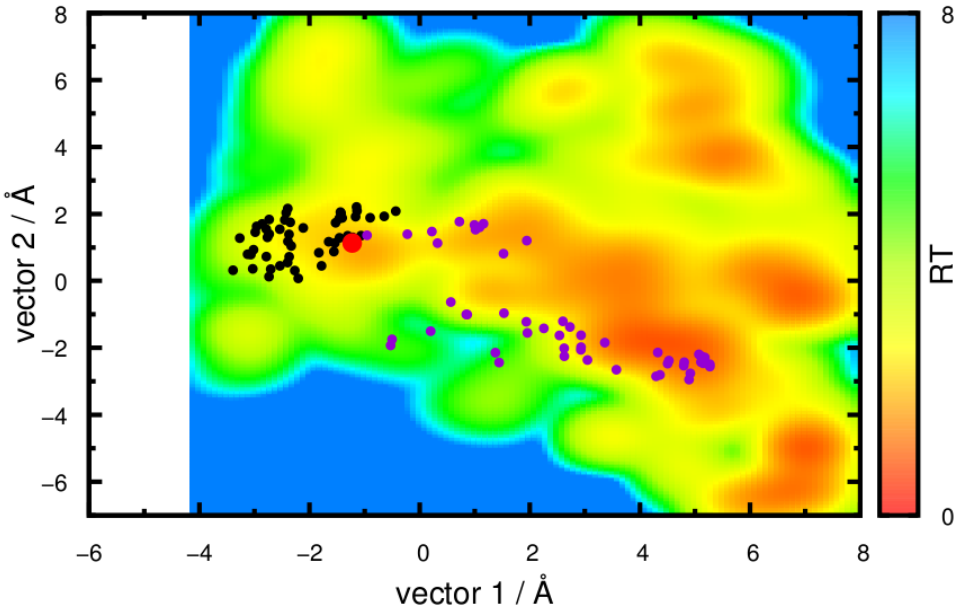


(a) pH 2

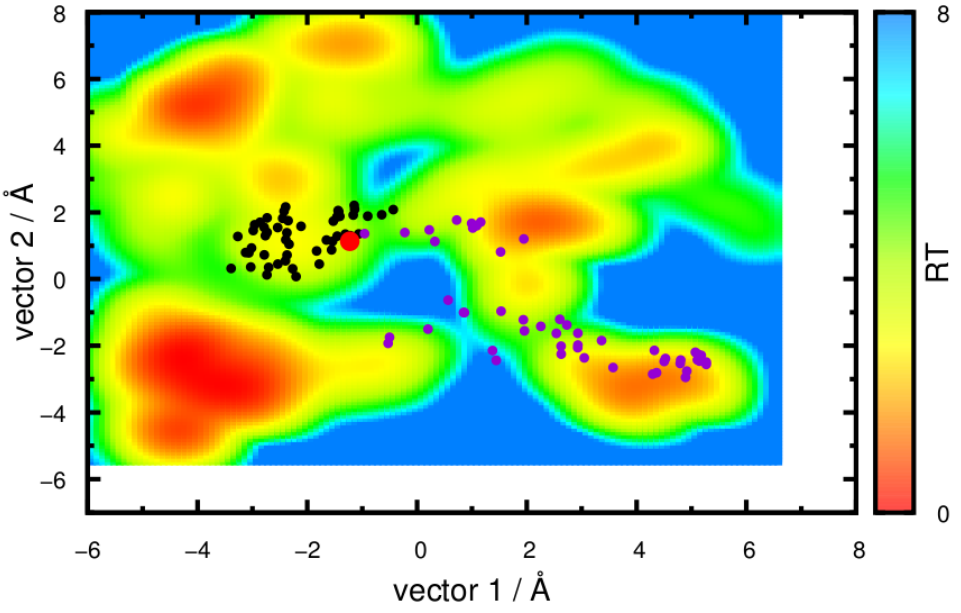


(b) pH 7

Figure 3.20: Simulation: 0.5c. Caption in **Figure 3.25**.

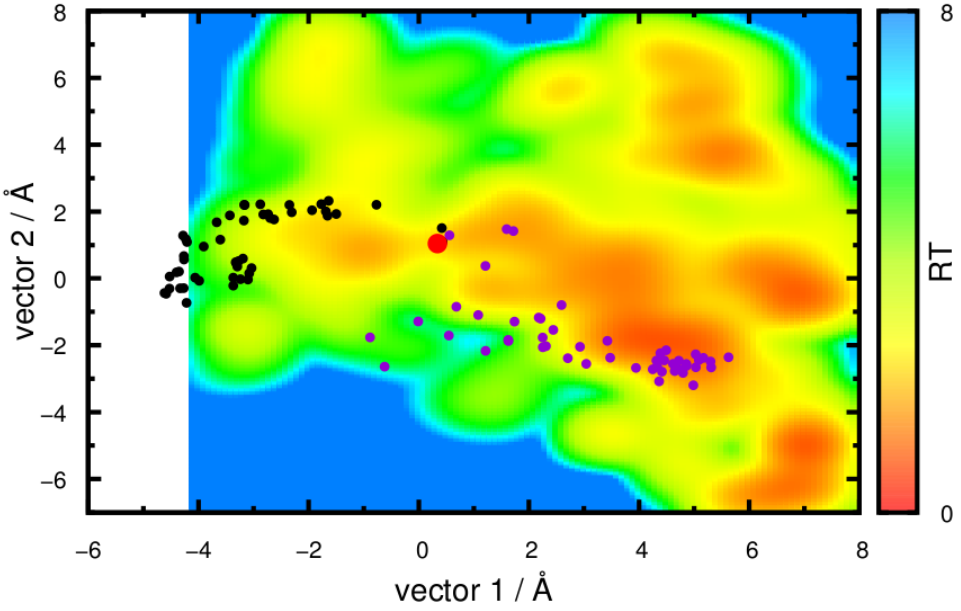


(a) pH 2

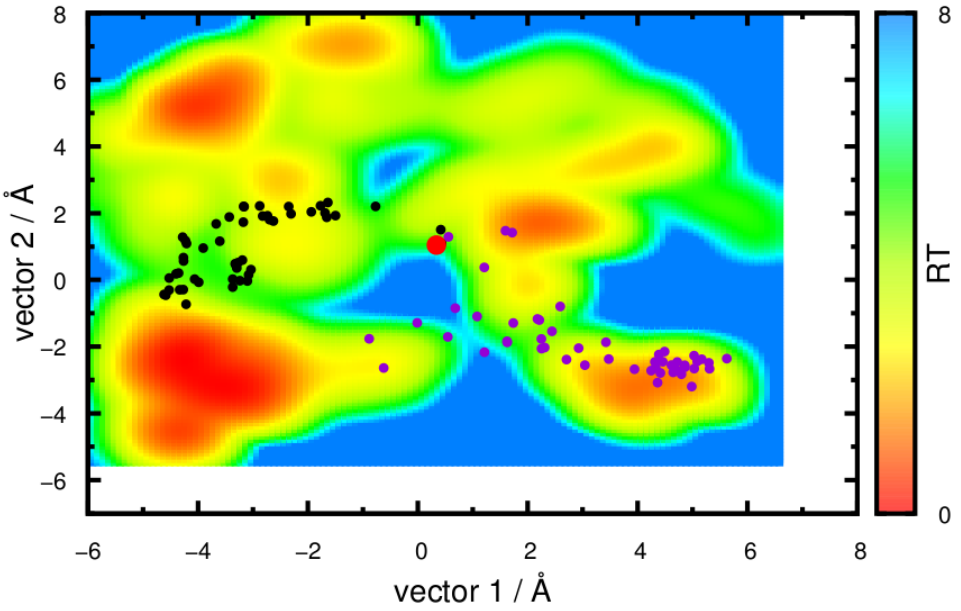


(b) pH 7

Figure 3.21: Simulation: 1.5a. Caption in **Figure 3.25**.

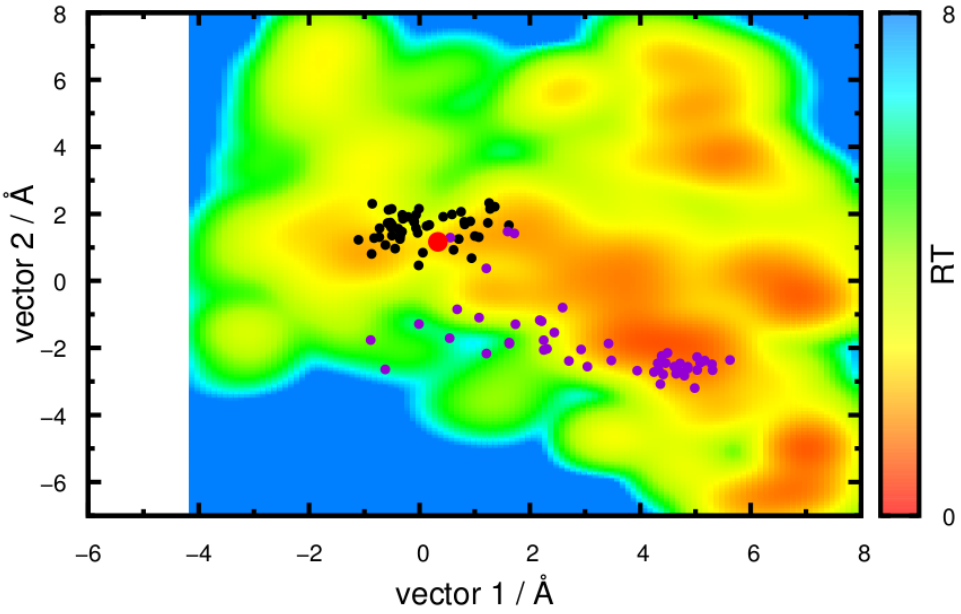


(a) pH 2

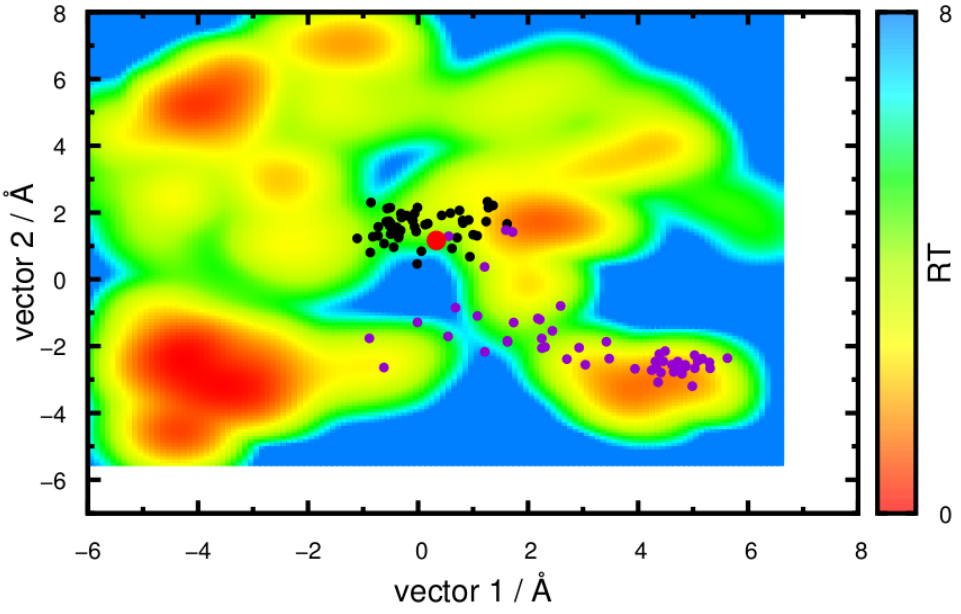


(b) pH 7

Figure 3.22: Simulation: 3.0a. Caption in **Figure 3.25**.

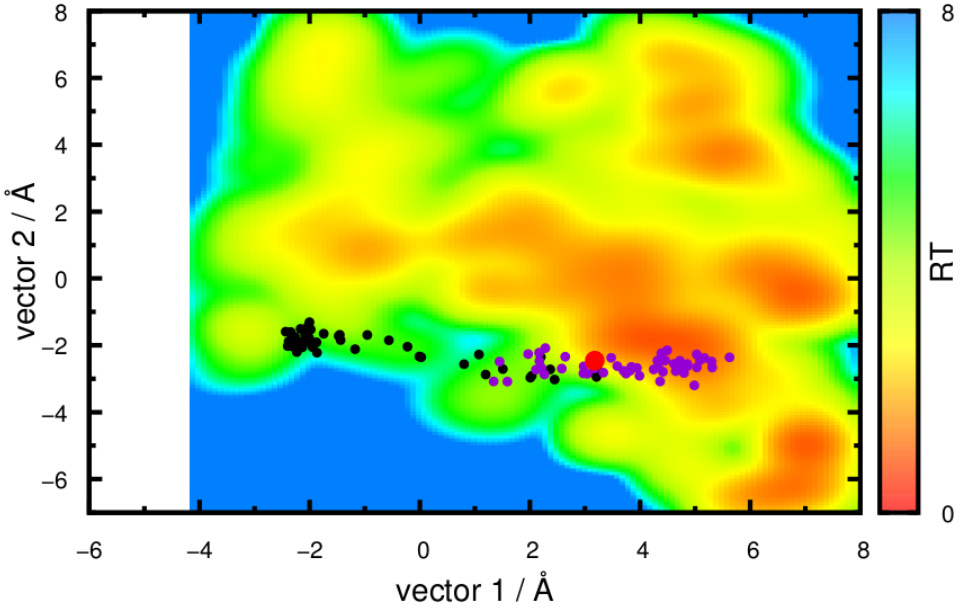


(a) pH 2

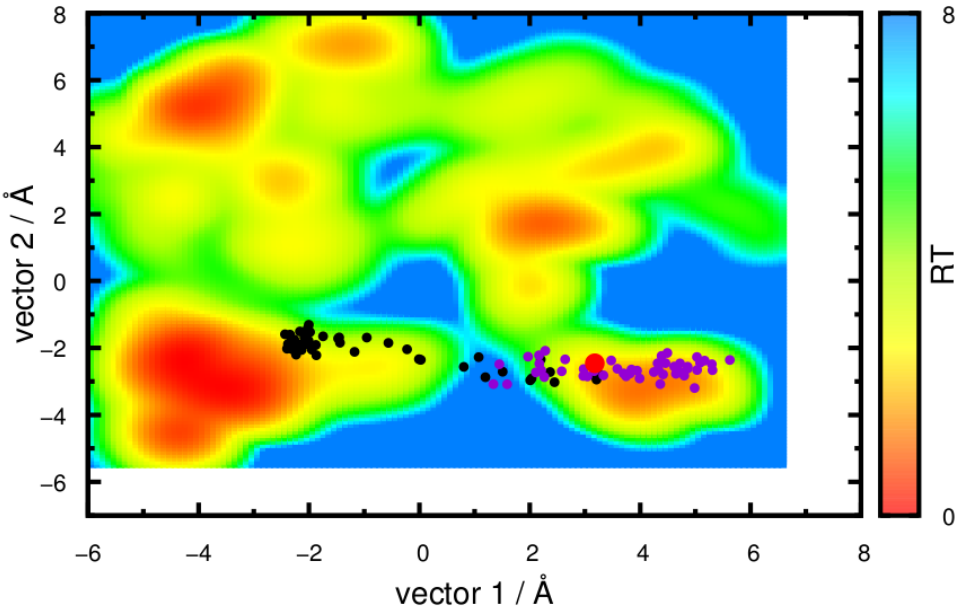


(b) pH 7

Figure 3.23: Simulation: 3.0c. Caption in **Figure 3.25**.

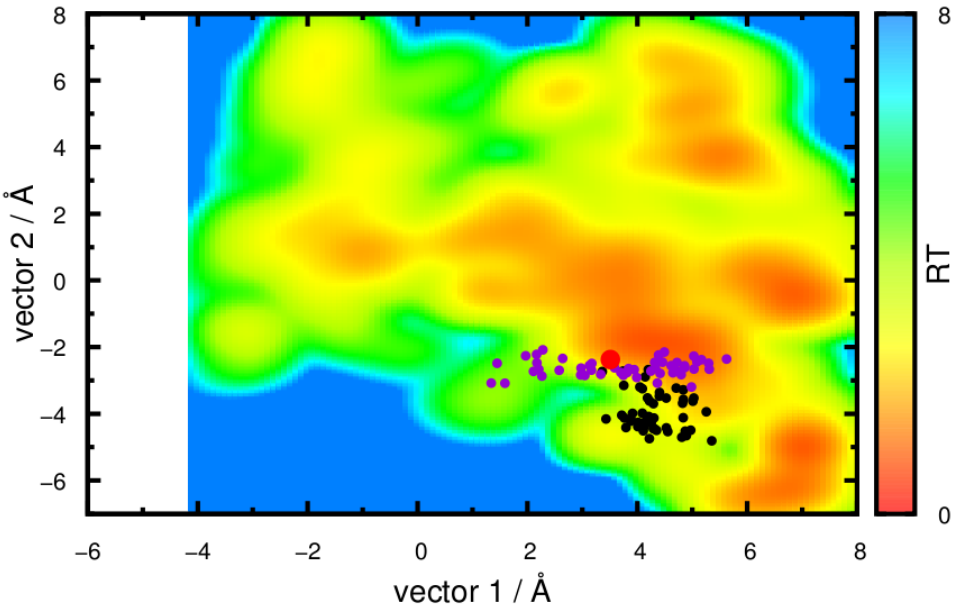


(a) pH 2

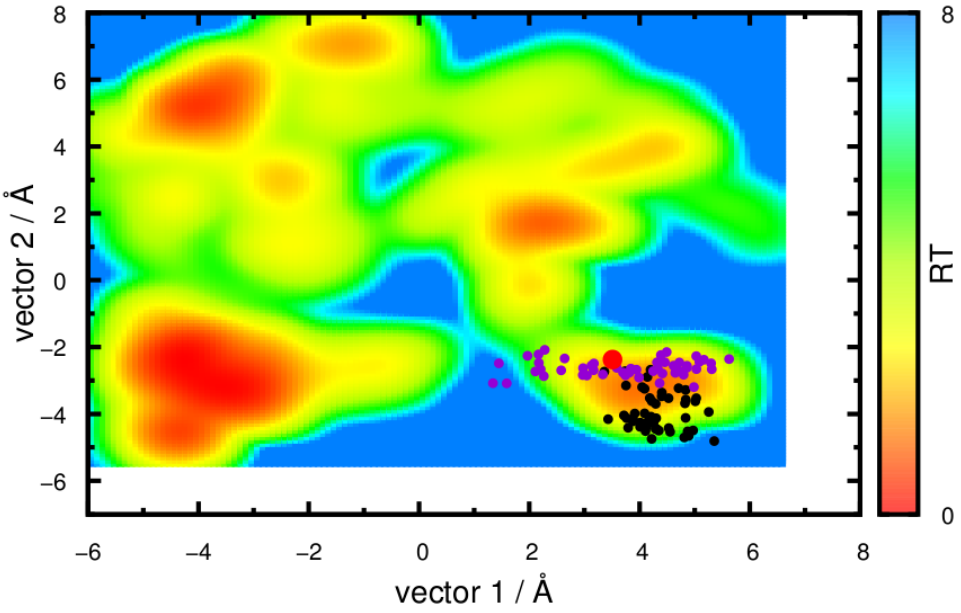


(b) pH 7

Figure 3.24: Simulation: 8.0a. Caption in **Figure 3.25**.



(a) pH 2



(b) pH 7

Figure 3.25: Simulation: 8.0b. Energy landscapes at pH 2 and pH 7 using the first two PC's and projections of short simulations (●) and equivalent snapshots of the pH2mis simulation (●). Starting points of short simulations are also marked (●). Simulations and pH of the landscapes are marked in captions.

- **0.5a - 0.5c** (**Figures E.1 - E.3** and **3.20**) - These simulations were started from an early time and they quickly trapped down into the deep basin of the landscape at pH 7. The protein conformation was still very close to the pH 7 fold and this situation was not altered in these simulations.

- **1.0a - 2.0c** and **3.0b** (**Figures E.4 - E.12**, **3.21** and **E.14**) - In this group of simulations, the pH 7 seems to have a large stabilizing effect. After the change in the pH the protein goes down to a basin near to the starting point and it stays there until the end of the simulation.

- **3.0a** (**Figures 3.22** and **E.13**) - In this simulation the protein leaves a typical zone of pH 2 and goes towards the pH 7 deeper basin. Despite this simulation never reaches the bottom of the basin it comes very close to it. Therefore, this indicates that a partial reversibility event occurred in this simulation.

- **3.0c**, **5.0c** and **8.0b** (**Figures 3.23**, **E.15**, **E.18**, **E.20** and **3.25**) - These simulations do not seem to go in the direction of a basin. They spread on the landscape without an evident trend. However, they spread on a direction different from the one taken by the pH2mis simulation.

- **5.0a**, **5.0b** and **8.0c** (**Figures E.16**, **E.17** and **E.21**) - The reversibility event is also suggested by these projections. Here, the protein leaves a typical zone of pH 2 and goes in the direction of the pH 7 basin.

- **8.0a** (**Figures E.19** and **3.24**) - According to the PCA, this simulation shows the most extensive reversibility transition. In this simulation the protein starts in a typical zone of pH 2 far from the deepest basin of pH 7 and it ends in the border of this basin, crossing a large distance in the landscape.

The result obtained for the simulation started at $t = 8.0$ ns (replicate a) led us to evaluate in detail the structural behavior of this replicate. In **Figure**

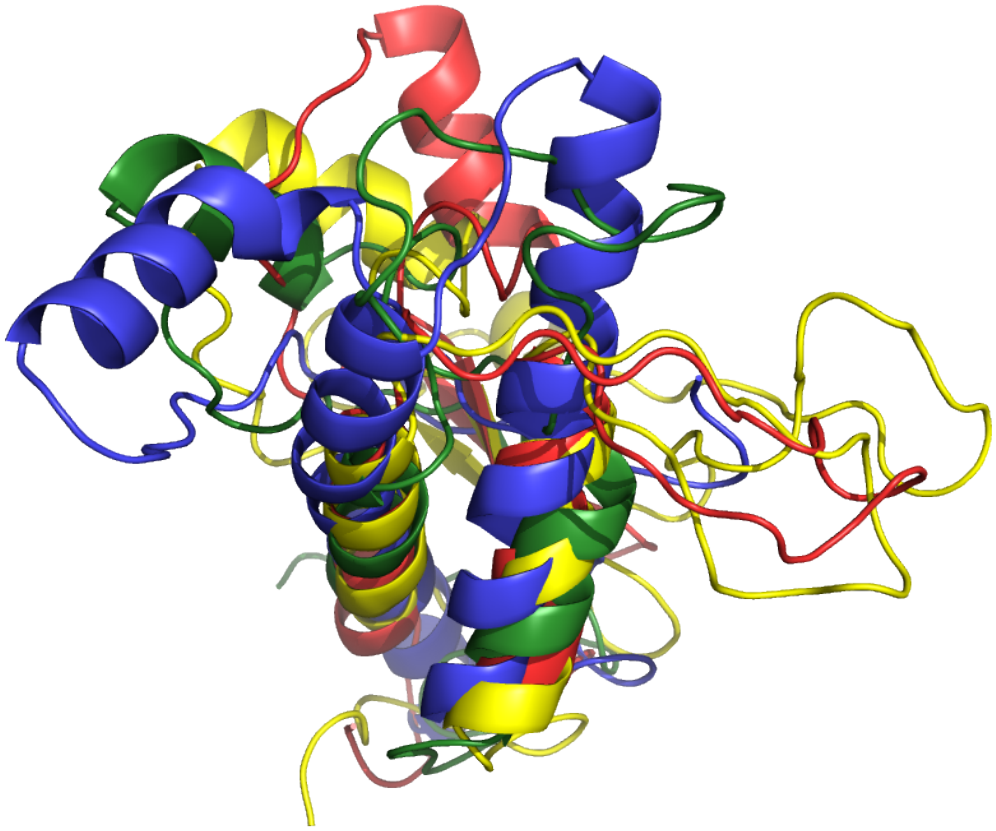


Figure 3.26: Conformational transition observed after changing the pH to 7 in a short simulation (starting point: 8.0 ns, replicate: a) ■ Starting snapshot; ■ t = 3.0 ns; ■ Final snapshot (t = 10.0 ns). ■ Starting snapshot of pH2mis and pH7mis simulations. Residues of the HB and HC helices were fitted to the reference structure (Starting snapshot of pH2mis and pH7mis simulations).

3.26 it is represented the starting, an intermediate and the final snapshots of this simulation and the starting snapshot of the pH2mis and pH7mis simulations. The figure identifies the two main regions of the protein that are changing in the direction of the reference structure. The loop between HB and HC is moving in the direction of the elongated arrangement of these two helices in the original structure. Also, HA, maintaining its overall fold, is changing its position towards the reference conformation. Remarkably, these observations confirm that the pH change in this particular replicate originated, in 10 ns, an almost complete reversion of the effects caused by 8 ns of simulation at pH 2.

The PCA analysis proved to be a very useful tool to evaluate this type of transitions. This may be due to its ability to look at the overall fold of the protein and not to particular aspects like secondary structure. We observed different key behaviors in the obtained landscapes:

- In most simulations, pH 7 had a stabilizing effect and the protein populated a zone near its starting point.
- In three simulations the protein spread over the landscape.
- In five simulations, pH 7 clearly reverts the effects caused by pH 2 on the protein.

In a general overview, it is very difficult to conclude undoubtedly about the reversibility of the observed transition. Nevertheless, we only observed reversibility events until the eighth nanosecond of the original simulation. This also indicates that the most determinant effects in the misfolding transition occurred early in the pH2mis simulation, rendering the reversibility event difficult to observe in the long simulations.

3.6 Force Field Comparison - 53a6 vs 43a1

Although the main goal of this project was to evaluate if the misfolding transition of PrP was reversible, the large loss of helical structure in our simulations at pH 7 led us to investigate the effect of a different force field. There are studies suggesting that long simulations with the 53a6 force field destabilise the helices in peptides [**79**] and even in proteins [**78**]. Therefore, if this problem is affecting our system, the force field may be masking the observation of the reversibility event in our simulations (specially the long ones, which are more prone to force field deficiencies). We decided to check this by running our long simulations (originally performed with the 53a6 force field) with the 43a1 force field. We analyzed the results in two different ways: we calculated the average values of helix and β-structure content in the 20-40 ns segment of the simulations (**Table 3.1**) and the variation of these secondary structure elements after the first 10 ns of simulation (**Table 3.2**). The helix structure was computed according with the DSSP criterion.

Table 3.1 shows that in most of the simulations the average value of helix content is higher in simulations performed with the 43a1 than in the ones performed with 53a6. In terms of β-structure, in 7 simulations the value is higher in the 43a1 and in other 5 simulations the opposite happens. These results suggest that the 53a6 force field destabilises helix structures when compared with the 43a1 force field. In the β-structures there is no evidence of over or under stabilization between these two force fields. The results were obtained by comparing the equilibrated segments (20-40 ns) of the simulations and suggest that a significant part of the helix loss observed in our simulations might come from a force field problem. This helix instability can oppose important conformational transitions essential to observe reversibility events. We also tested the difference between the two force fields in short simulations (**Table 3.2**). The results in this table do not show a significant difference between the two force fields. This indicates that the helix destabilization effect is only relevant in long equilibrium simulations. These results assure that our short simulations are not affected by

Starting	Helices			β-structure		
time	53a6	43a1	Difference	53a6	43a1	Difference
11.0 ns	17.24	31.56	**-14.32**	16.00	9.90	6.10
21.0 ns	17.45	17.91	**-0.46**	24.66	17.65	7.02
29.0 ns	1.52	13.64	**-12.12**	26.69	29.63	**-2.94**
30.0 ns	8.93	12.02	**-3.09**	28.43	31.39	**-2.95**
40.0 ns	9.64	12.33	**-2.69**	25.19	32.90	**-7.71**
41.2 ns	8.17	14.42	**-6.25**	28.30	28.51	**-0.21**
42.0 ns	11.27	13.71	**-2.44**	33.72	29.27	4.45
47.5 ns	12.51	10.90	1.61	18.99	23.01	**-4.03**
49.0 ns	13.52	13.68	**-0.17**	21.09	18.41	2.68
54.5 ns	10.49	11.85	**-1.37**	23.08	24.28	**-1.20**
55.7 ns	11.20	11.31	**-0.10**	21.17	24.79	**-3.62**
60.0 ns	9.07	16.70	**-7.62**	19.89	16.65	3.24

Table 3.1: Average values of helix and β-structure content in the equilibrated segment (20 - 40 ns) of long simulations with the 43a1 and 53a6 force fields and difference between the values for the two force fields. **Bold highlight** indicates when the value is higher in 43a1 force field.

this destabilization effect present in 53a6 force field and, therefore, guarantee that in the first 10 ns of simulations, we are looking at a real pH induced conformational transition.

Despite the observed effects in long simulations with 53a6 force field, these effects should not be too drastic. In the previous work, Campos et al [**1**] performed many long simulations and, at pH 7, the PrP showed a stable folded structure. Moreover, a clear trend of helix and β structures variation with pH was observed which is completely independent of any force field destabilization effect.

Starting	Helices			β-structure		
time	53a6	43a1	Difference	53a6	43a1	Difference
11.0 ns	-0.7	-1.2	0.5	-0.2	-0.1	**-0.1**
21.0 ns	0.0	-0.3	0.3	1.2	1.1	0.1
29.0 ns	-0.7	-0.1	**-0.6**	1.0	0.1	0.9
30.0 ns	0.0	-0.4	0.4	-0.2	0.5	**-0.7**
40.0 ns	-0.4	-0.3	**-0.1**	0.1	0.8	**-0.7**
41.2 ns	-0.8	-0.2	**-0.6**	-1.0	-0.8	**-0.2**
42.0 ns	0.2	0.3	**-0.1**	0.6	1.2	**-0.6**
47.5 ns	-0.1	-0.4	0.3	0.4	0.6	**-0.2**
49.0 ns	-0.2	-0.2	0.0	0.0	0.0	0.0
54.5 ns	0.6	-0.2	0.8	-0.2	0.3	**-0.5**
55.7 ns	0.2	-0.5	0.7	0.6	1.1	**-0.5**
60.0 ns	-0.6	0.4	**-1.0**	-0.2	0.3	**-0.5**

Table 3.2: Variation of helix and β-structure content after the first 10 ns of long simulations with the 43a1 and 53a6 force fields and difference between the values for the two force fields. **Bold highlight** indicates when the value is higher in 43a1 force field.

Chapter 4

Concluding Remarks

The main purpose of the present work was to evaluate if the misfolding transition observed in PrP at pH 2 by Campos et al. [**1**] is reversible. This problem was address by performing constant-pH MD simulations at pH 7 started from different transient conformations of the simulation in which an evident misfolding transition occurred. Two different approaches were used: long simulations were performed in order to try to observe complete reversibility phenomena and short simulations to look at the initial effects of the change to pH 7 on the protein. Our results were analyzed with several tools and approached from different points of view.

The DSSP and the N-O contacts tools allowed us to conclude that, after the change in pH, the new β-structures and long range contacts that are formed, can be transient and meta-stable in many cases or very stable in another. These stable β structures at pH 7 might also difficult the observation of an evident reversibility phenomenon. The N-O contacts tool also allowed to conclude that pH 7 had a stabilizing effect on short range contacts. This stabilizing effect is also evident in other analysis.

The N-O contacts tool revealed to be less sensitive than DSSP to small alterations in the protein conformation. The results with this tool also supported

the conclusion that pH 7 has a stabilizing effect in PrP's conformations.

Neither the DSSP criterion nor the N-O contacts tool allowed us to identify evident reversibility transitions. The use of these tools to evaluate reversibility events is limited because they depend too much on the secondary structure recovery, which might be problematic due to the long timescales needed for such events to take place.

In order to complement the previous approaches, we used four other methods which look at the global fold of the protein. The radius of gyration and SAS do not allow to conclude much about the reversibility of the misfolding of PrP. Nevertheless, we observed that this two properties were correlated when the protein exhibited a considerable number of structural contacts. However, this correlation was lost in simulations started in later times in which PrP had lost most of its original contacts. These properties also supported the fact that most important and non-recovered transitions happened quite early in the original simulation at pH 2. These transitions were always characterized by increased radius of gyration and SAS.

The most significant information about the reversibility of the misfolding of PrP was obtained with last two types of analysis: RMSD and PCA. There are two main conclusions obtained from these results: pH 7 has a large local stabilizing effect in the conformations of PrP and reversibility is only observed if we start our pH 7 simulations from early conformations of the pH2mis simulation, those where no major conformational transition took place. Indeed, in a significant number of simulations, the reversion of the protein misfolding caused by the pH 2 was evident.

Table 4.1 summarizes the results obtained with the RMSD against a central structure (RMSD) (section 3.4), difference of RMSD average against pH2mis and pH7mis simulations (RMSD - dif) (section 3.4) and PCA (section 3.5). In this table, most fields indicate that pH 7 had a stabilizing effect in the protein. Campos et al. have previously observed this effect [**1**]. In 11 simulations there is at least one methodology indicating a reversibility event. In 3 of them this event was suggested by the three methods (3.0a,

Simulation	RMSD	RMSD - dif	PCA
0.5a	-	X	-
0.5b	√	-	-
0.5b	-	-	-
1.0a	-	-	-
1.0b	-	-	-
1.0c	-	-	-
1.5a	√	√	-
1.5b	√	√	-
1.5c	-	-	-
2.0a	√	√	-
2.0b	√	√	-
2.0c	-	-	-
3.0a	√	√	√
3.0b	-	-	-
3.0c	-	X	X
5.0a	√	-	√
5.0b	-	-	√
5.0c	-	-	X
8.0a	√	√	√
8.0b	X	√	X
8.0c	√	√	√

Table 4.1: Summary table of the results obtained with the RMSD against a central structure (RMSD), difference of RMSD average against pH2mis and pH7mis simulations (RMSD - dif) and PCA. √ means that the reversibility event is suggested by the property, X means the opposite, and "-" means that the pH 7 had a stabilizing or some non-conclusive effect in the protein.

8.0a and 8.0c). Only 4 simulations suggested that the transition was irreversible and never by the three methodologies. Moreover, there is only one simulation (8.0b) where the results are contradictory, which suggests that the use of these three measures together can be a valuable tool to study these kind of complex fold transitions in proteins.

The PCA allowed us to obtain two energy landscapes (at pH 2 and at pH 7). A correct energy landscape can help us to conclude if the stabilization of the folded state is thermodynamic or kinetic. However, in the obtained landscapes it was not possible to know the relative depth of the basins at pH 2 and at pH 7. To know that, it would be necessary to observe many transitions between the several basins obtained. Therefore, with these simulations it was not possible to conclude about the thermodynamic / kinetic control of the folded state.

We also tested if the 53a6 force field had a destabilizing effect in PrP secondary structure. It was possible to conclude that the 53a6 force field destabilizes the helix structures in long simulations when compared with the 43a1 force field, but not in the short 10 ns ones. Similar observations have already been reported in the literature for peptides [**79**] and proteins [**78**].

In the long simulations performed, we did not observe significant reversibility events. The force field may be hindering the occurrence of these events in terms of secondary structure. Nevertheless, even with a perfect force field, the reversibility of such altered systems might only occur in a large computationally inaccessible timescale.

The short simulations proved to be very useful to evaluate the initial effect of changing pH to 7. These simulations showed large stabilizing effects induced by pH 7 and, even, reversion of the initial conformational changes induced by pH 2. The effects of the force field are negligible in these short simulations.

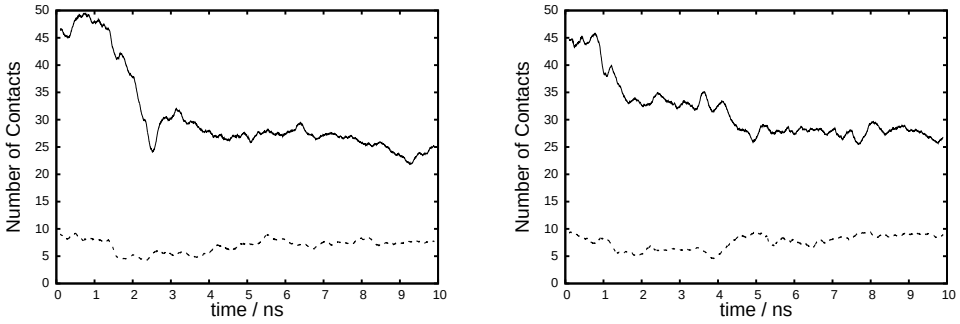
The reversibility of the misfolding in PrP could not be completely clarified

with this study mainly because we were not able to sample a full transition from a misfolded structure to a correctly folded one. Therefore, we can not state categorically that this transition is reversible or not. Surprisingly, we can attest on the relatively fast fold correction of PrP when its low pH-induced misfolding is not too drastic. Among such misfolded transitions that can be promptly recovered with an increase of pH, we can find an interesting long movement of helix HA, that has already been proposed to be a consequence of protonation and that could also be observed in PrP mutations associated with Creutzfeld-Jacob disease [**2**].

As an outlook, we think that similar studies could be done in a systematic way to study the effect of the change to pH 7 in all other replicates of the previous work [**1**]. This would give a more complete idea of the tendency to revert the effects induced by pH 2. Also, a detailed study of long simulations using other force fields could help us minimize the general force field limitations and could increase our chances of observing a complete reversibility transition.

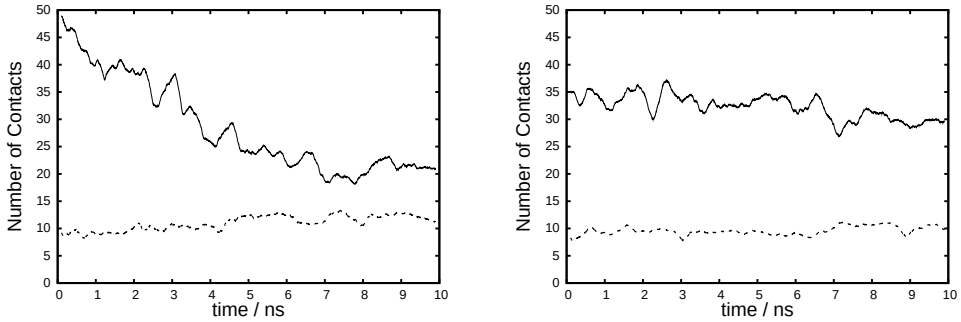
Appendix A

Number of important contacts maintained in short simulations



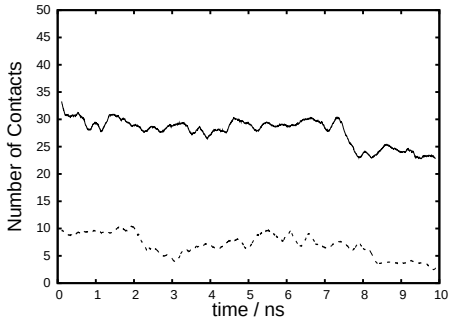
(a) 0.5 ns(a) **(b)** 0.5 ns(b)

Figure A.1: See caption in **Figure A.7**.

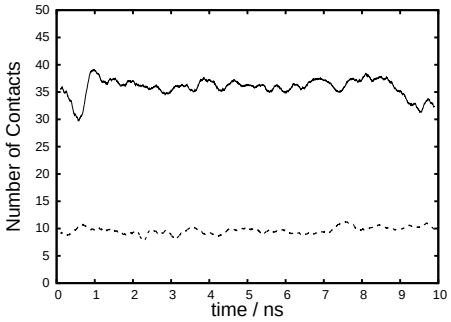


(a) 0.5 ns(c)

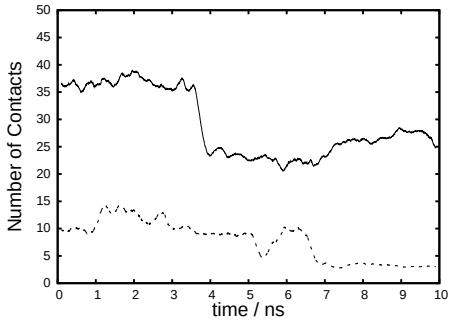
(b) 1.0 ns(a)



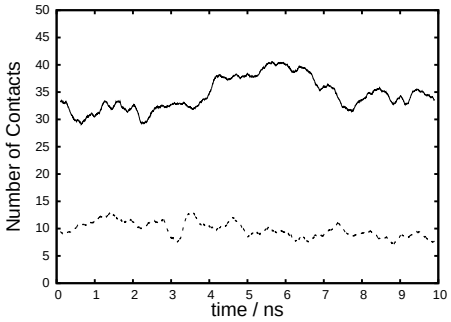
(c) 1.0 ns(b)



(d) 1.0 ns(c)

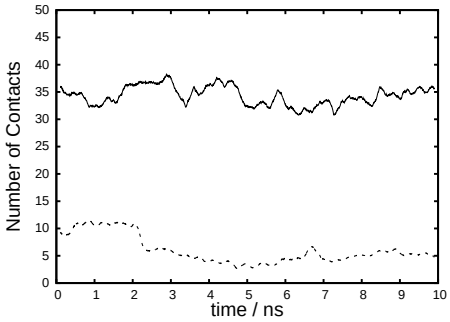


(e) 1.5 ns(a)

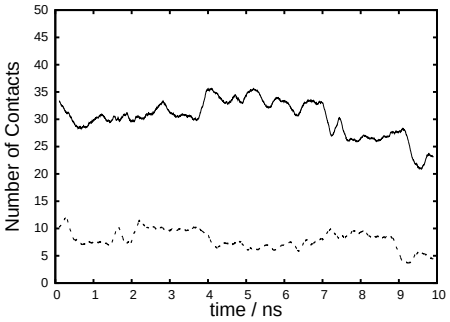


(f) 1.5 ns(b)

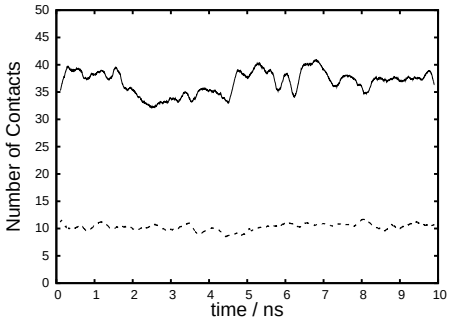
Figure A.2: See caption in **Figure A.7**.



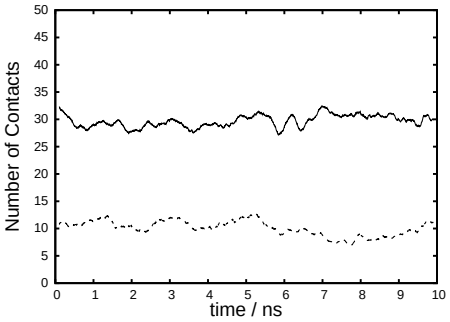
(a) 1.5 ns(c)



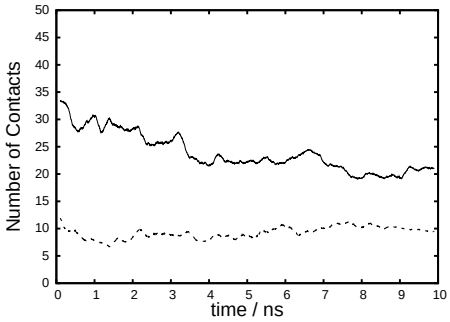
(b) 2.0 ns(a)



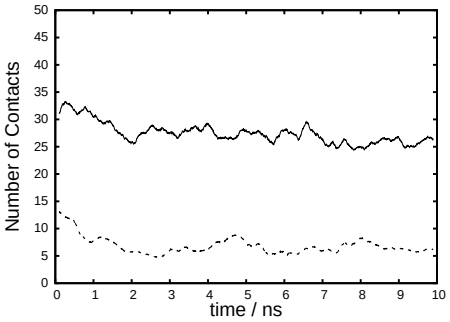
(c) 2.0 ns(b)



(d) 2.0 ns(c)

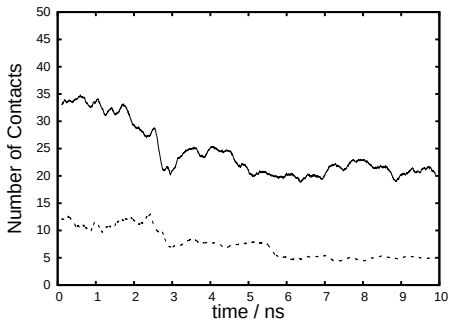


(e) 3.0 ns(a)

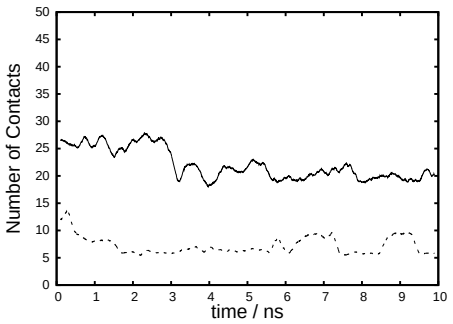


(f) 3.0 ns(b)

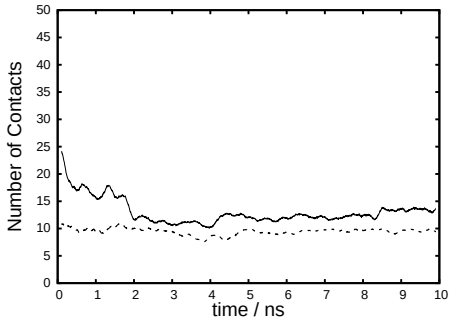
Figure A.3: See caption in **Figure A.7**.



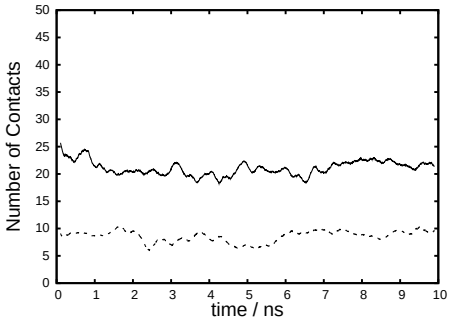
(a) 3.0 ns(c)



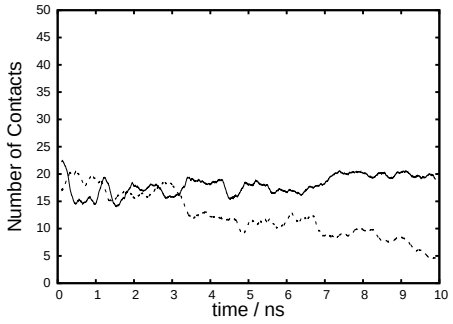
(b) 5.0 ns(a)



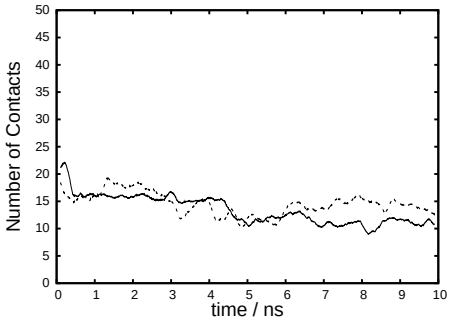
(c) 5.0 ns(b)



(d) 5.0 ns(c)

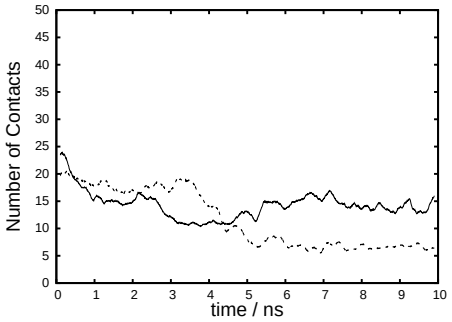


(e) 8.0 ns(a)

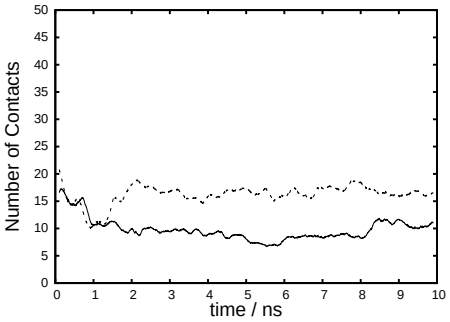


(f) 8.0 ns(b)

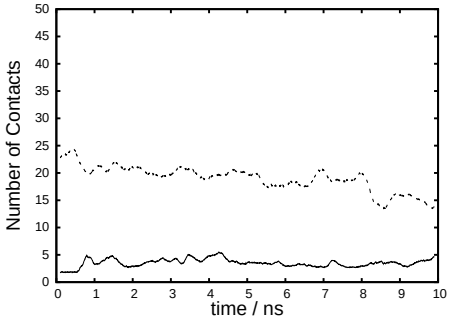
Figure A.4: See caption in **Figure A.7**.



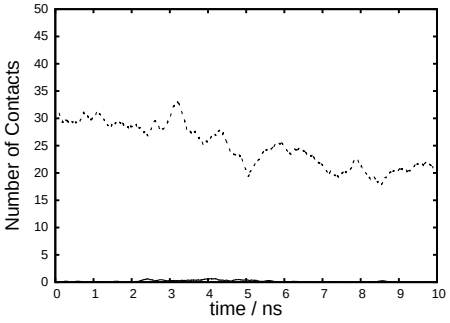
(a) 8.0 ns(c)



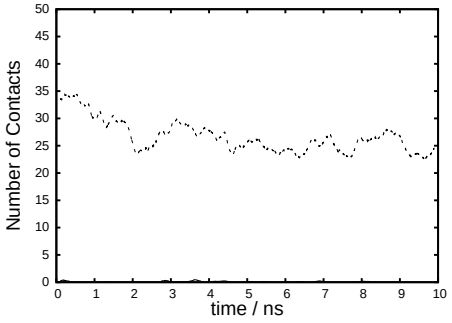
(b) 11.0 ns



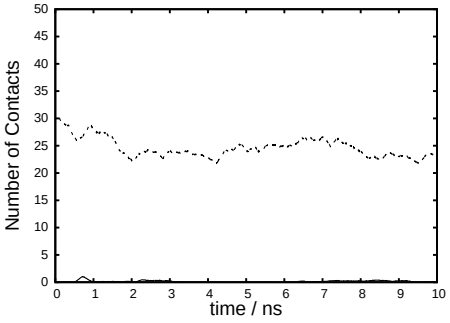
(c) 21.0 ns



(d) 29.0 ns

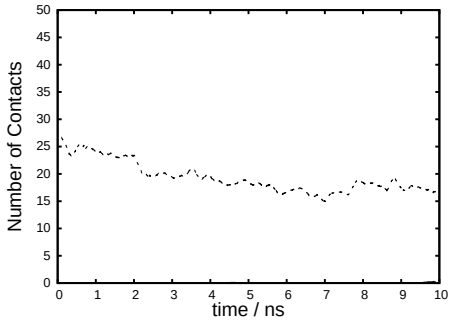


(e) 30.0 ns

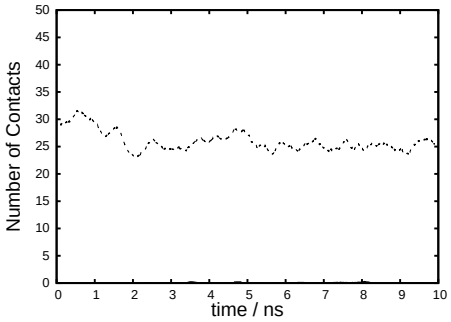


(f) 40.0 ns

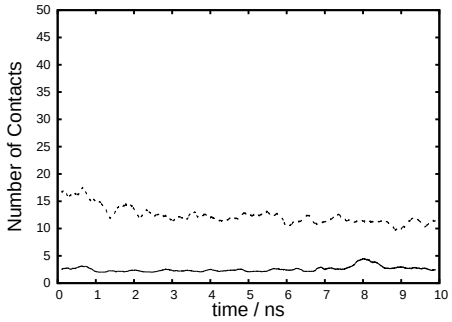
Figure A.5: See caption in **Figure A.7**.



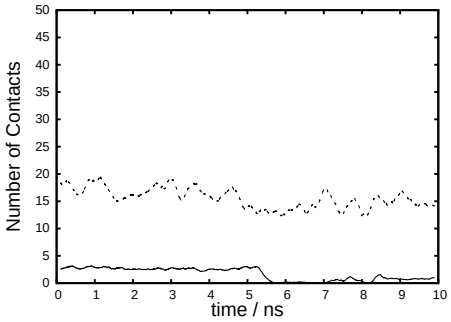
(a) 41.2 ns



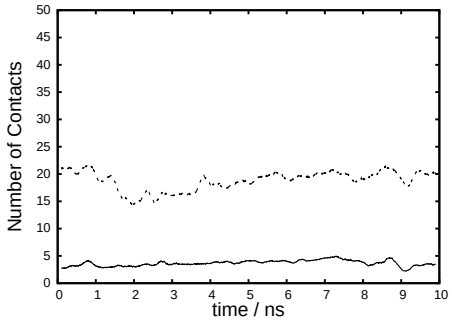
(b) 42.0 ns



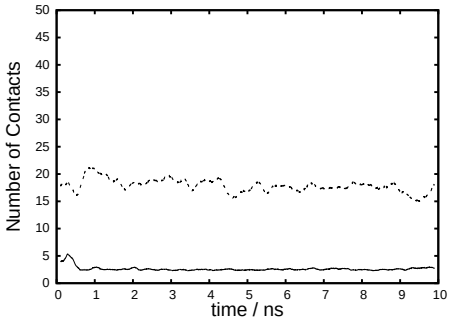
(c) 47.5 ns



(d) 49.0 ns

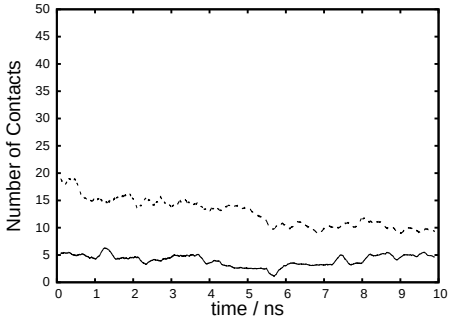


(e) 54.5 ns



(f) 55.7 ns

Figure A.6: See caption in **Figure A.7**.

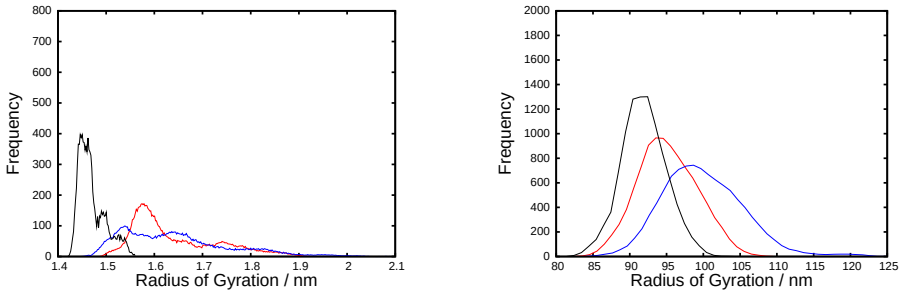


(a) 60.0 ns

Figure A.7: Variation of the number of important contacts maintained in short simulations. imp-pH7 (—), imp-pH2 (- -). Starting points of short simulations are marked in captions.

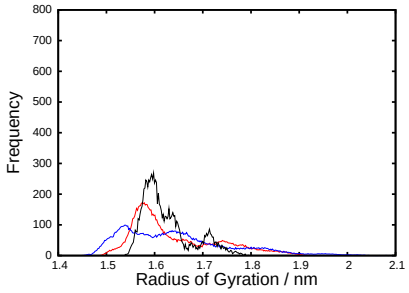
Appendix B

Histograms of Radius of Gyration and SAS of Long Simulations

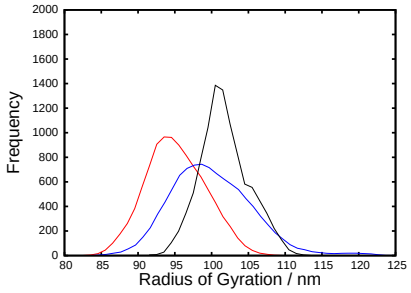


(a) Radius of Gyration **(b)** SAS

Figure B.1: Simulation: 11.0 ns. See caption in **Figure B.12**.

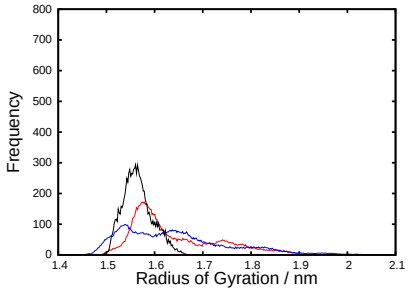


(a) Radius of Gyration

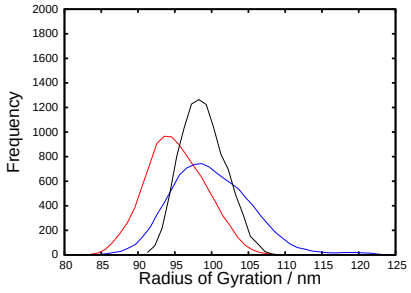


(b) SAS

Figure B.2: Simulation: 21.0 ns. See caption in **Figure B.12**.

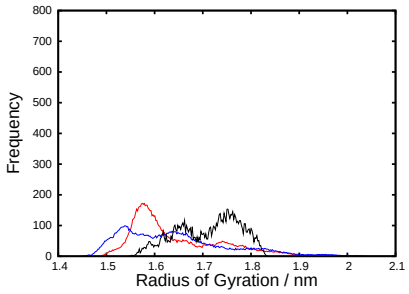


(a) Radius of Gyration

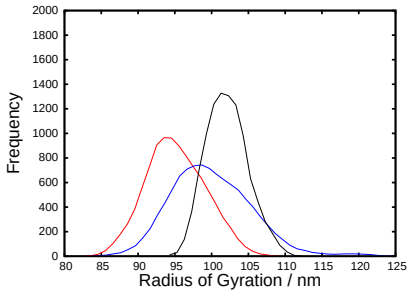


(b) SAS

Figure B.3: Simulation: 29.0 ns. See caption in **Figure B.12**.

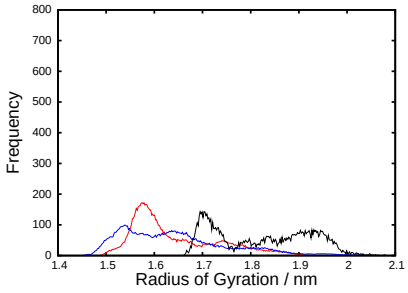


(a) Radius of Gyration

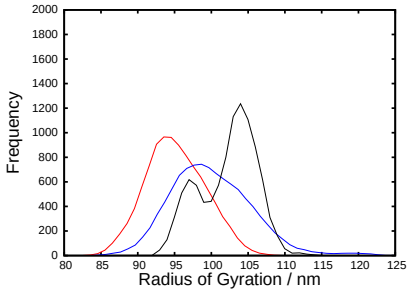


(b) SAS

Figure B.4: Simulation: 30.0 ns. See caption in **Figure B.12**.

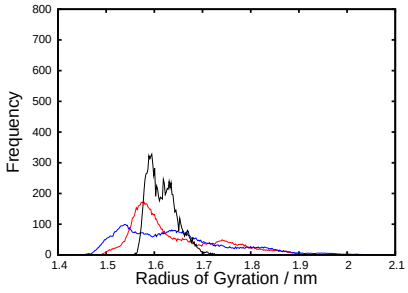


(a) Radius of Gyration

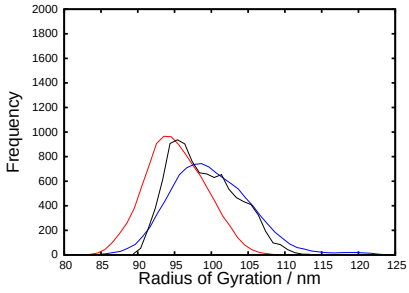


(b) SAS

Figure B.5: Simulation: 40.0 ns. See caption in **Figure B.12**.

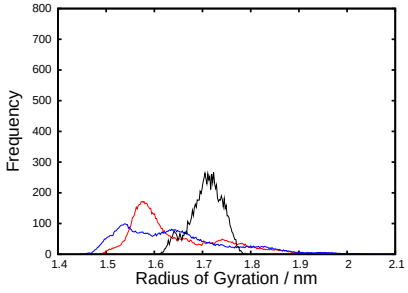


(a) Radius of Gyration

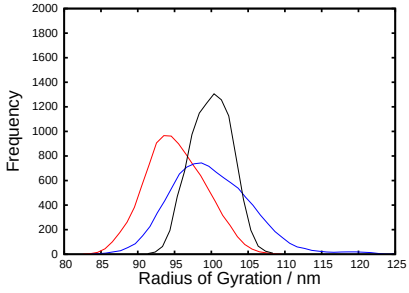


(b) SAS

Figure B.6: Simulation: 41.2 ns. See caption in **Figure B.12**.

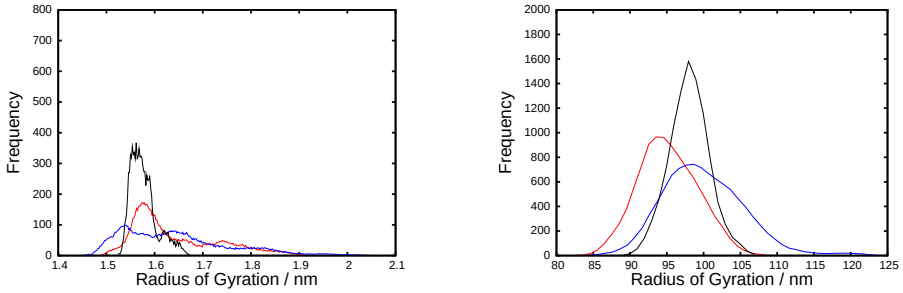


(a) Radius of Gyration



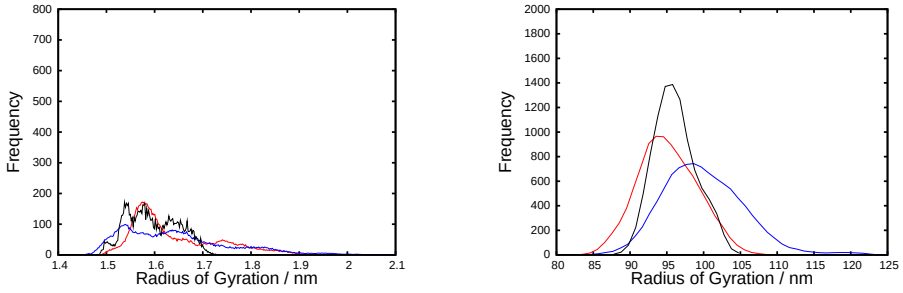
(b) SAS

Figure B.7: Simulation: 42.0 ns. See caption in **Figure B.12**.



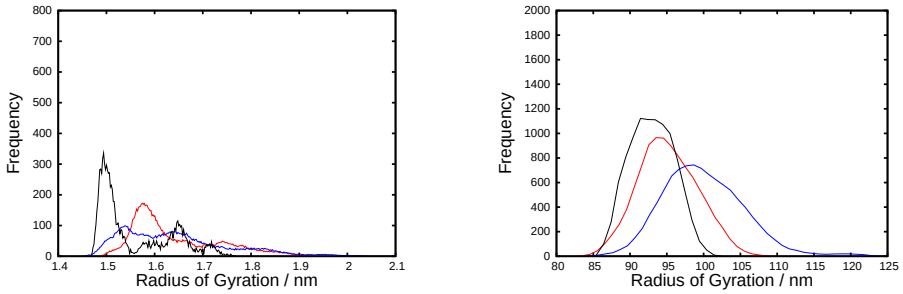
(a) Radius of Gyration **(b)** SAS

Figure B.8: Simulation: 47.5 ns. See caption in **Figure B.12**.



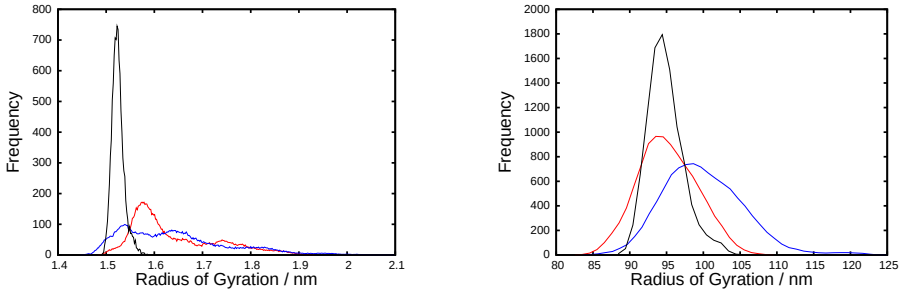
(a) Radius of Gyration **(b)** SAS

Figure B.9: Simulation: 49.0 ns. See caption in **Figure B.12**.



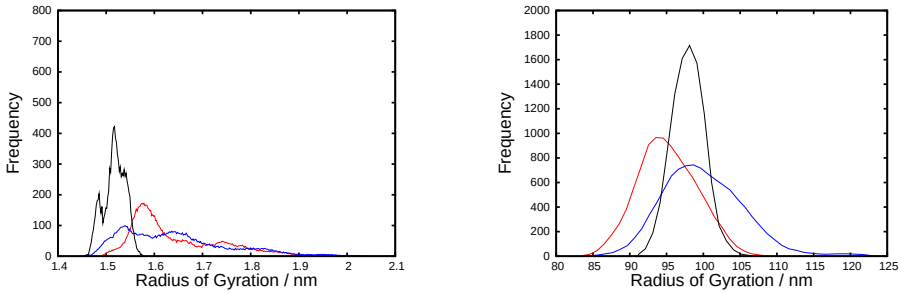
(a) Radius of Gyration **(b)** SAS

Figure B.10: Simulation: 54.5 ns. See caption in **Figure B.12**.



(a) Radius of Gyration **(b)** SAS

Figure B.11: Simulation: 55.7 ns. See caption in **Figure B.12**.



(a) Radius of Gyration **(b)** SAS

Figure B.12: Simulation: 60.0 ns. Histograms of Radius of Gyration and SAS. – Average values of pH 2 simulations – Average values of pH 7 simulations – Long simulations (starting points are marked in the captions).

Appendix C

RMSD plots of short simulations

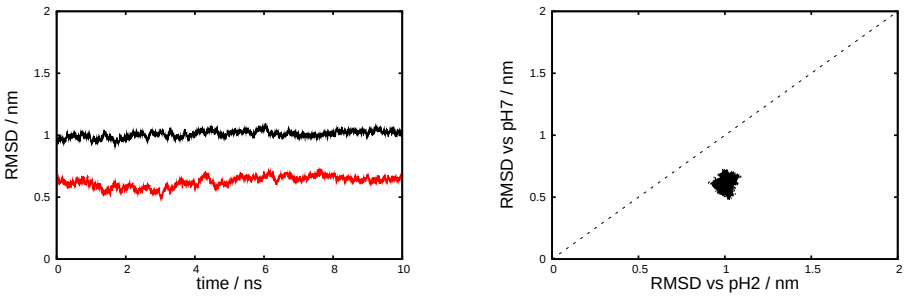


Figure C.1: Simulation: 0.5 ns(a). See caption in **Figure C.33**.

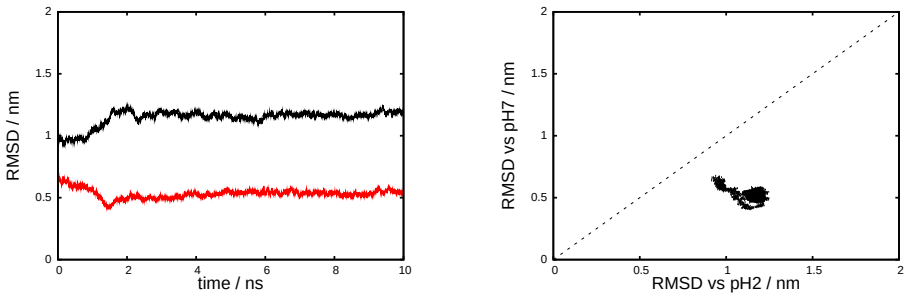


Figure C.2: Simulation: 0.5 ns(b). See caption in **Figure C.33**.

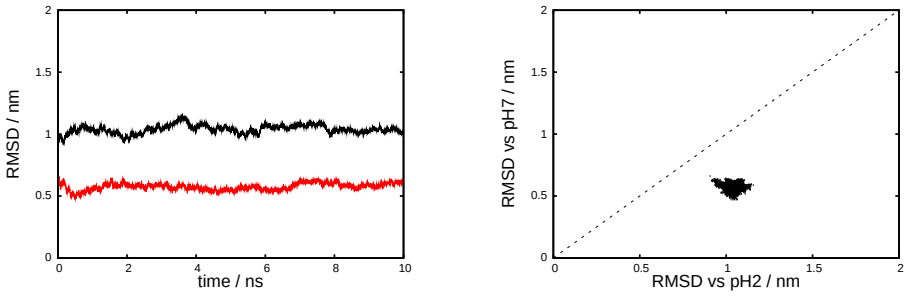


Figure C.3: Simulation: 0.5 ns(c). See caption in **Figure C.33**.

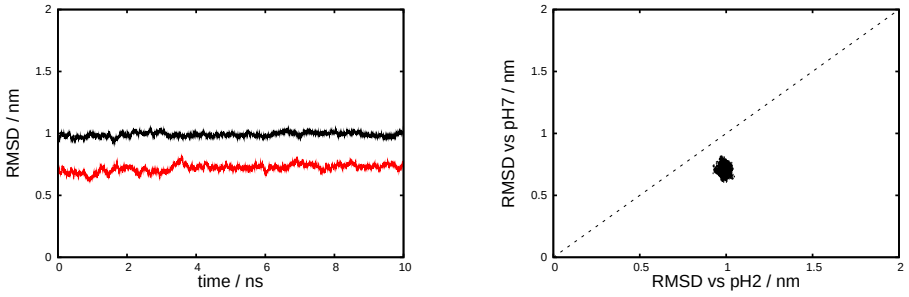


Figure C.4: Simulation: 1.0 ns(a). See caption in **Figure C.33**.

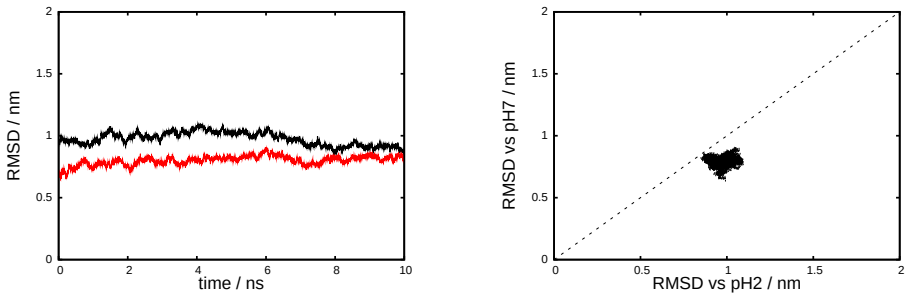


Figure C.5: Simulation: 1.0 ns(b). See caption in **Figure C.33**.

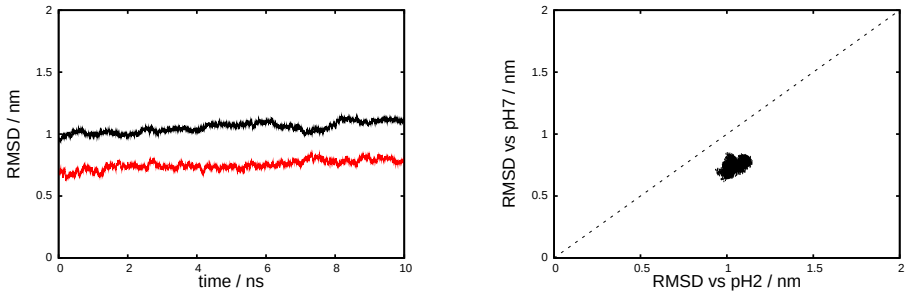


Figure C.6: Simulation: 1.0 ns(c). See caption in **Figure C.33**.

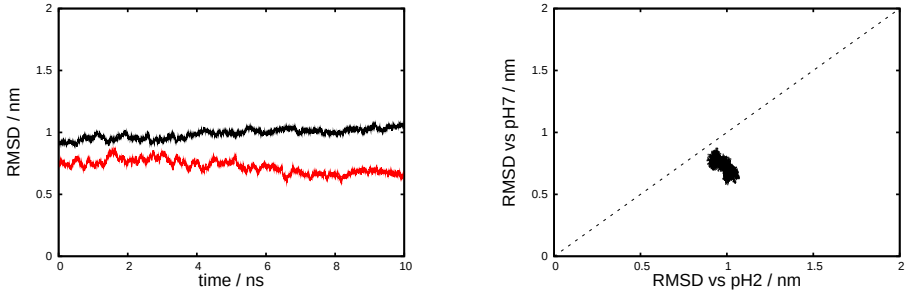


Figure C.7: Simulation: 1.5 ns(a). See caption in **Figure C.33**.

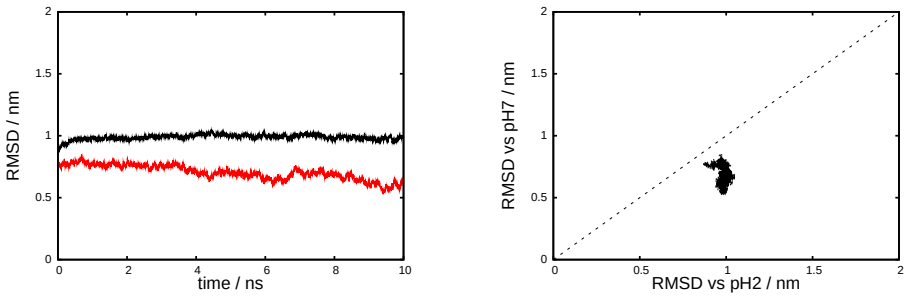


Figure C.8: Simulation: 1.5 ns(b). See caption in **Figure C.33**.

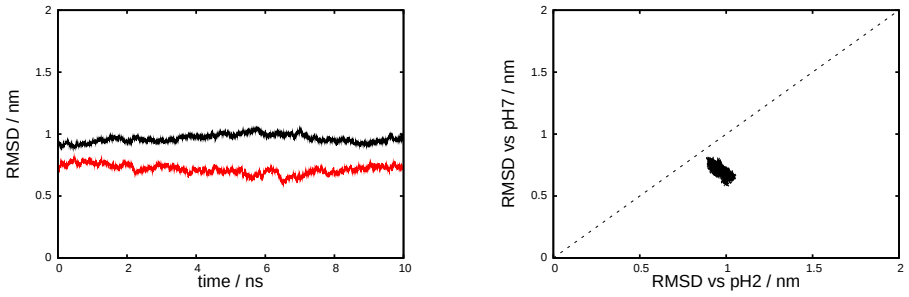


Figure C.9: Simulation: 1.5 ns(c). See caption in **Figure C.33**.

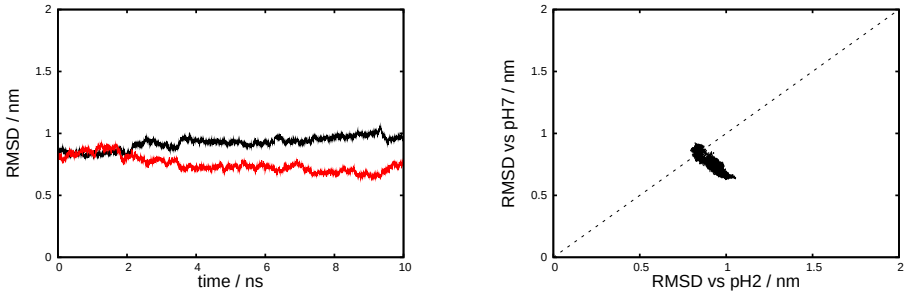


Figure C.10: Simulation: 2.0 ns(a). See caption in **Figure C.33**.

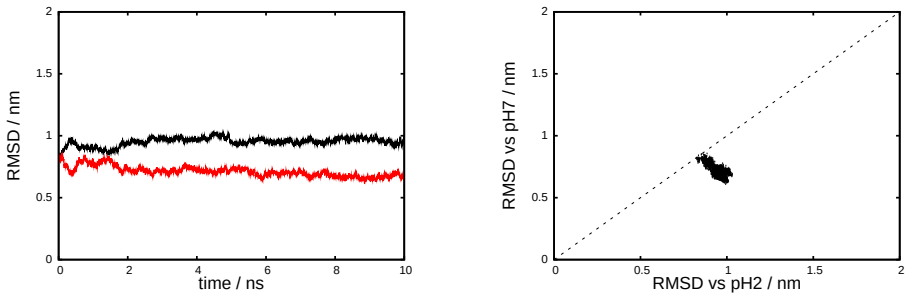


Figure C.11: Simulation: 2.0 ns(b). See caption in **Figure C.33**.

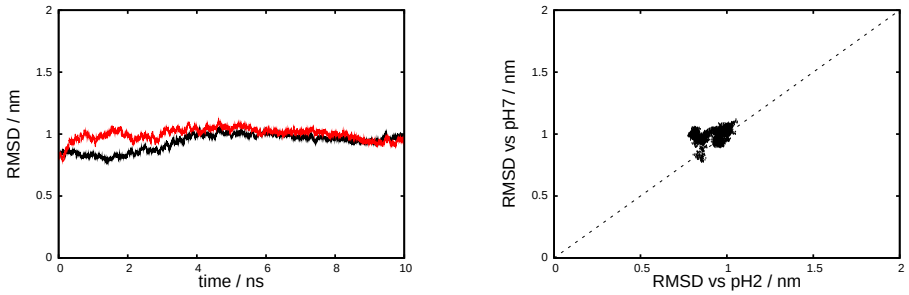


Figure C.12: Simulation: 2.0 ns(c). See caption in **Figure C.33**.

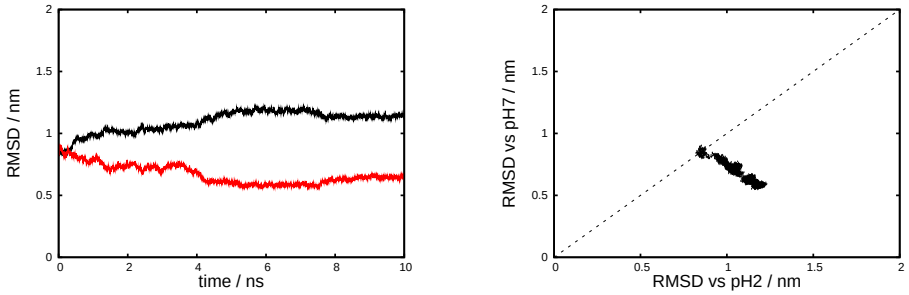


Figure C.13: Simulation: 3.0 ns(a). See caption in **Figure C.33**.

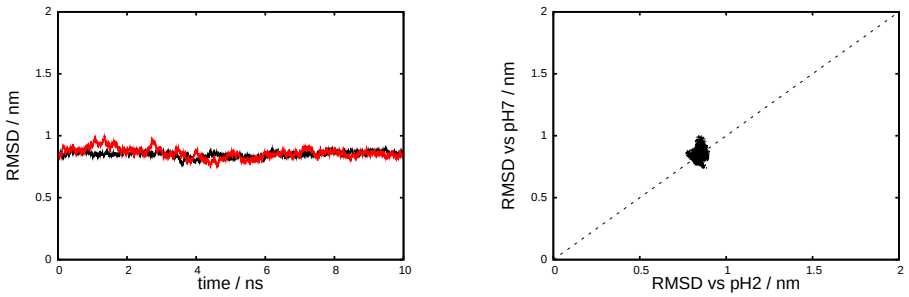


Figure C.14: Simulation: 3.0 ns(b). See caption in **Figure C.33**.

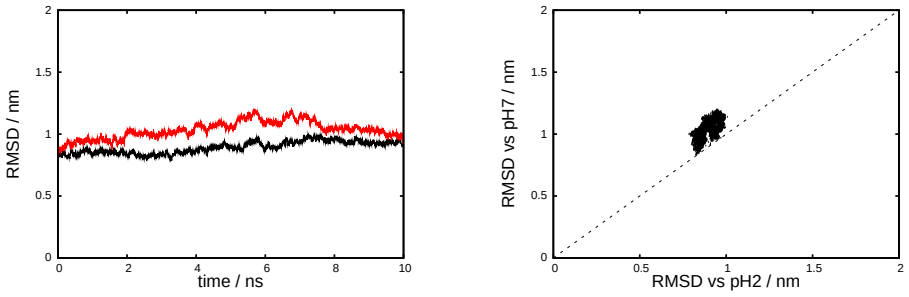


Figure C.15: Simulation: 3.0 ns(c). See caption in **Figure C.33**.

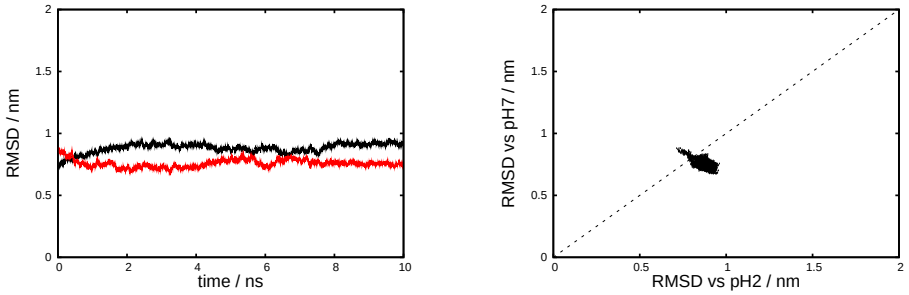


Figure C.16: Simulation: 5.0 ns(a). See caption in **Figure C.33**.

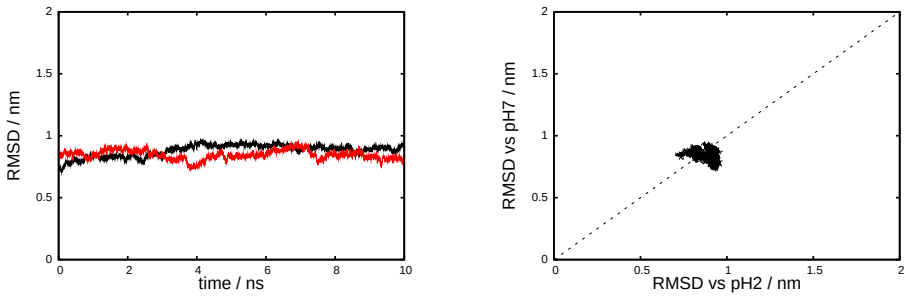


Figure C.17: Simulation: 5.0 ns(b). See caption in **Figure C.33**.

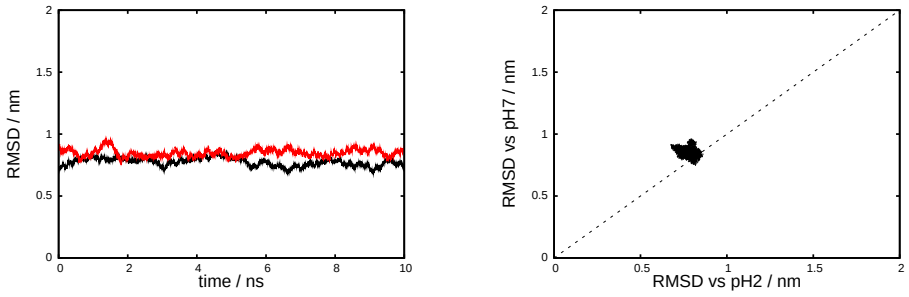


Figure C.18: Simulation: 5.0 ns(c). See caption in **Figure C.33**.

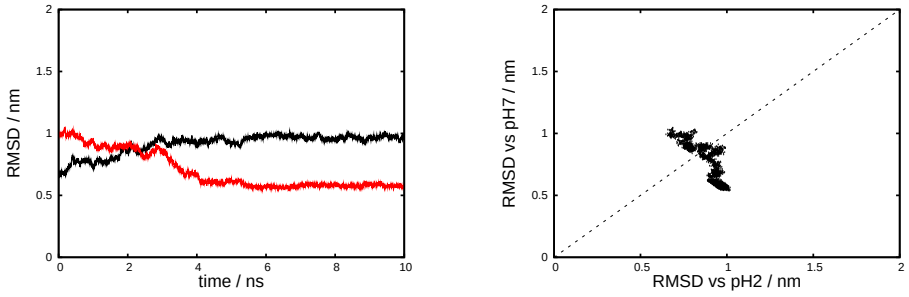


Figure C.19: Simulation: 8.0 ns(a). See caption in **Figure C.33**.

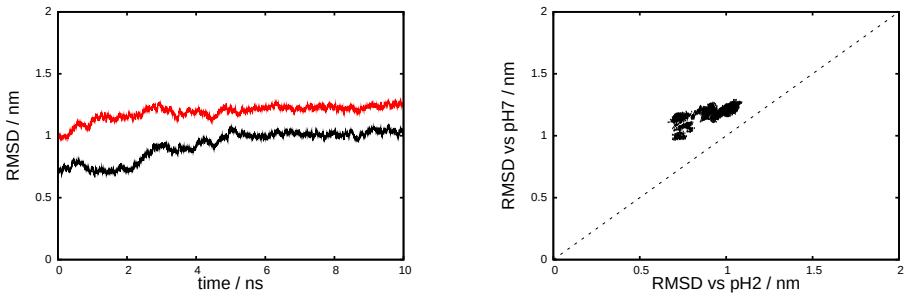


Figure C.20: Simulation: 8.0 ns(b). See caption in **Figure C.33**.

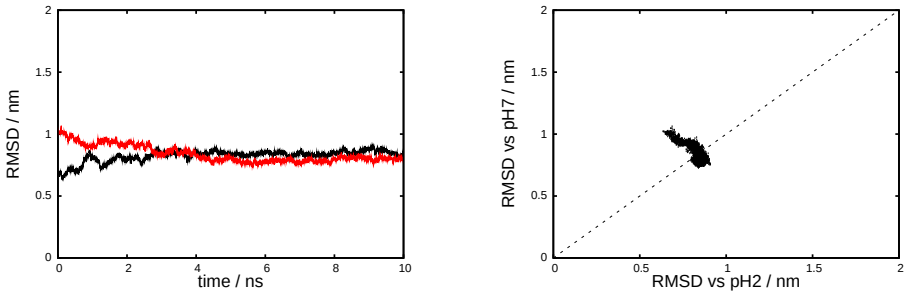


Figure C.21: Simulation: 8.0 ns(c). See caption in **Figure C.33**.

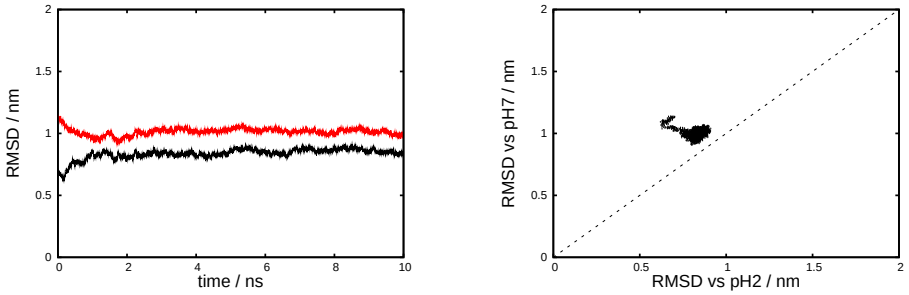


Figure C.22: Simulation: 11.0 ns. See caption in **Figure C.33**.

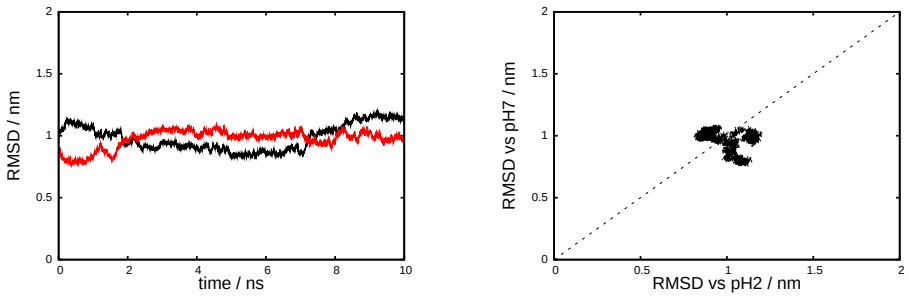


Figure C.23: Simulation: 21.0 ns. See caption in **Figure C.33**.

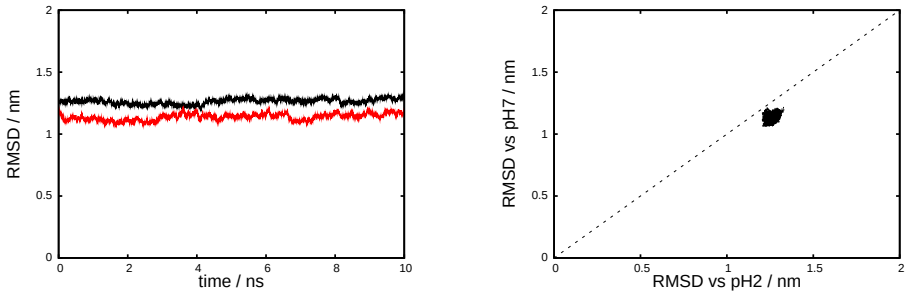


Figure C.24: Simulation: 29.0 ns. See caption in **Figure C.33**.

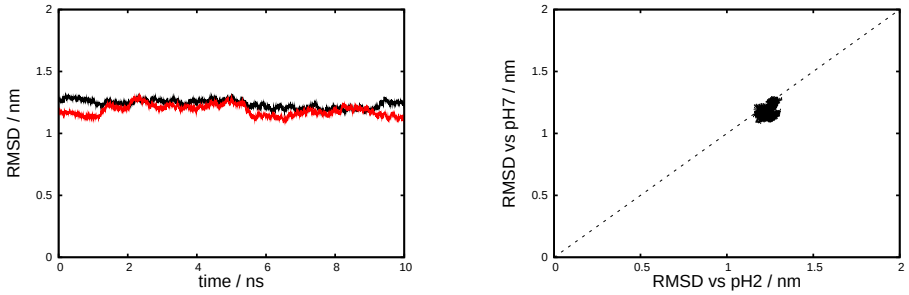


Figure C.25: Simulation: 30.0 ns. See caption in **Figure C.33**.

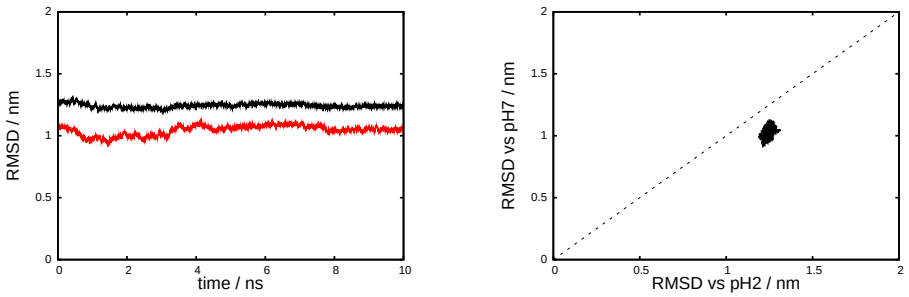


Figure C.26: Simulation: 40.0 ns. See caption in **Figure C.33**.

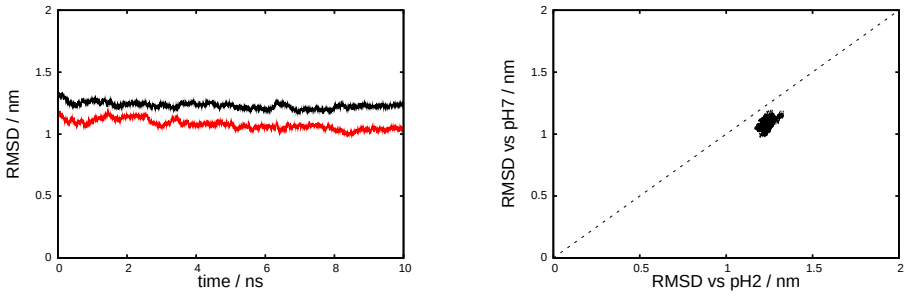


Figure C.27: Simulation: 41.2 ns. See caption in **Figure C.33**.

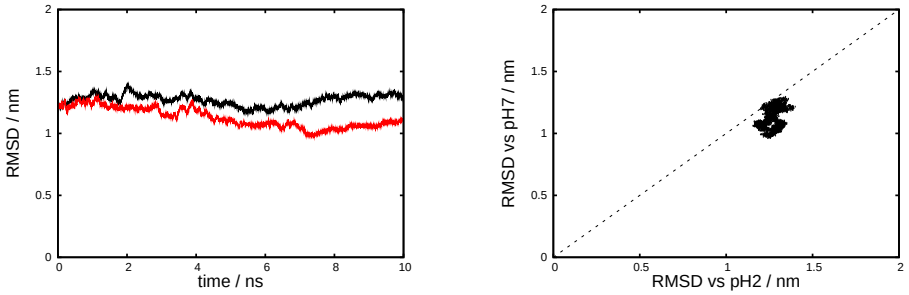


Figure C.28: Simulation: 42.0 ns. See caption in **Figure C.33**.

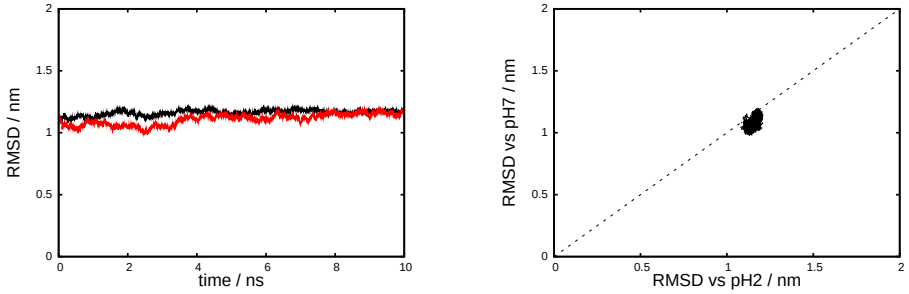


Figure C.29: Simulation: 47.5 ns. See caption in **Figure C.33**.

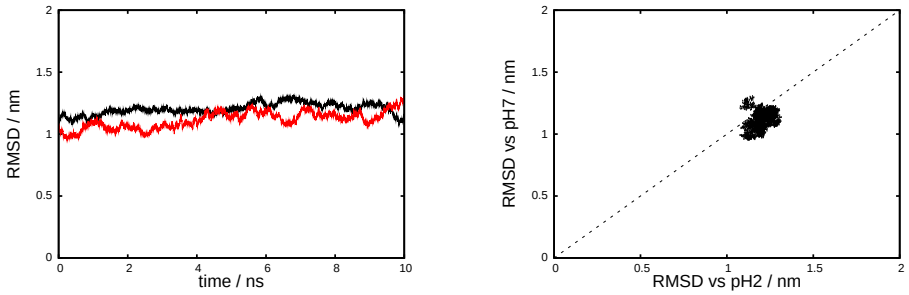


Figure C.30: Simulation: 49.0 ns. See caption in **Figure C.33**.

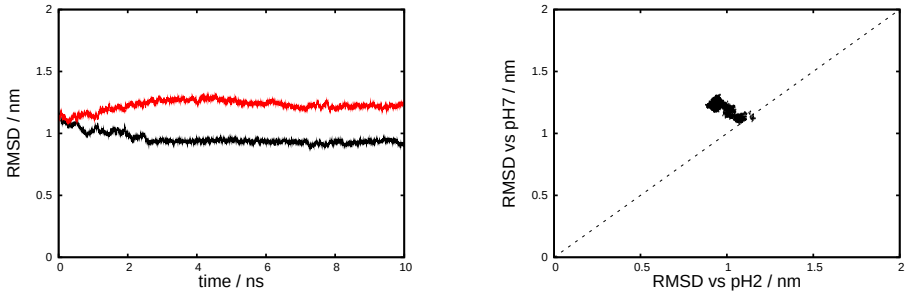


Figure C.31: Simulation: 54.5 ns. See caption in **Figure C.33**.

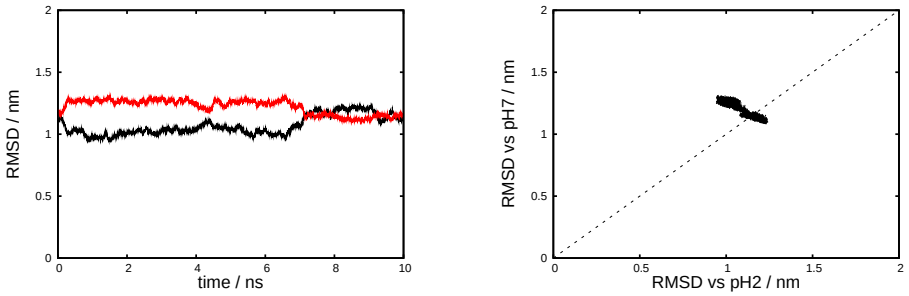


Figure C.32: Simulation: 55.7 ns. See caption in **Figure C.33**.

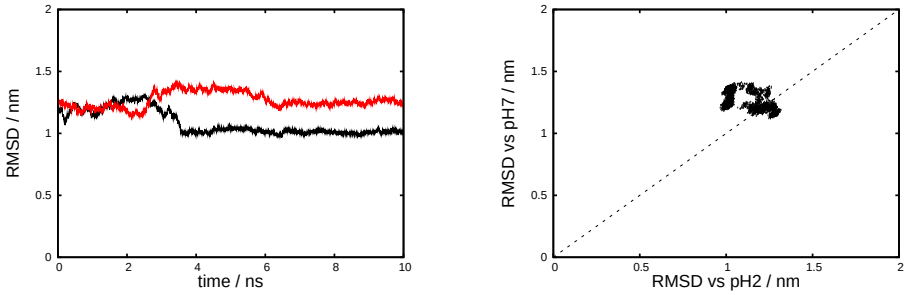
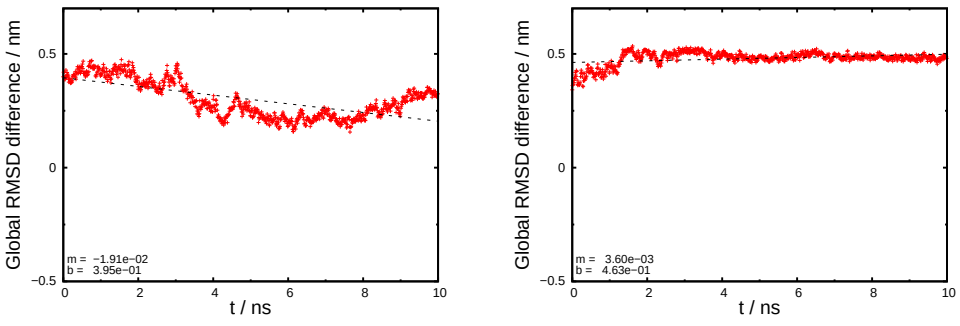


Figure C.33: Simulation: 60.0 ns. left) – RMSD vs average structure of pH 7 simulations, – RMSD vs average structure of pH 2 simulations. **right)**, • (RMSD vs average structure of pH 7 simulations) vs (RMSD vs average structure of pH 2 simulations), ·· x=y line.

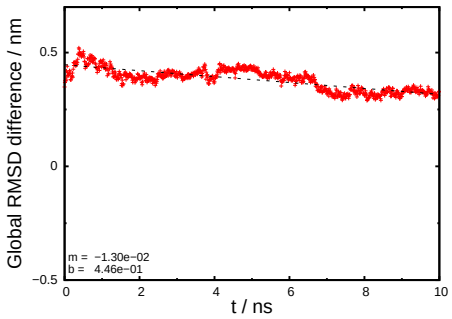
Appendix D

Difference of average RMSD value vs pH2mis and pH7mis conformations in the short simulations

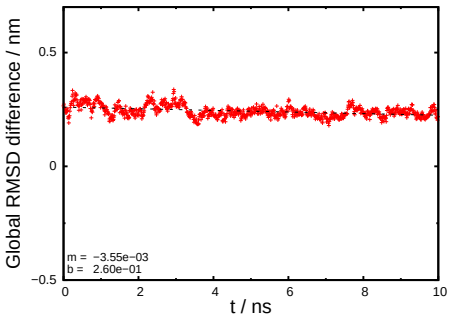


(a) 0.5 ns(a) **(b)** 0.5 ns(b)

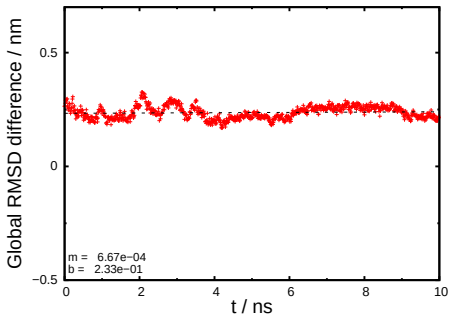
Figure D.1: See caption in **Figure D.5**.



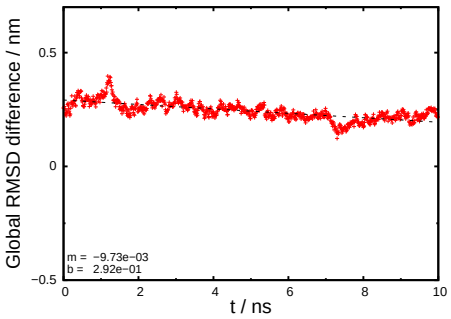
(a) 0.5 ns(c)



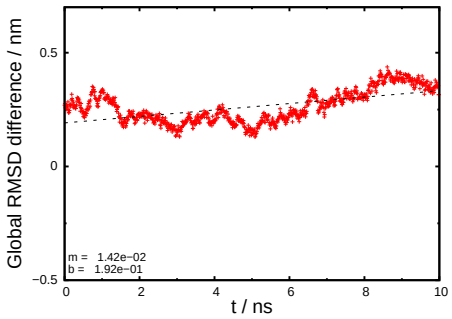
(b) 1.0 ns(a)



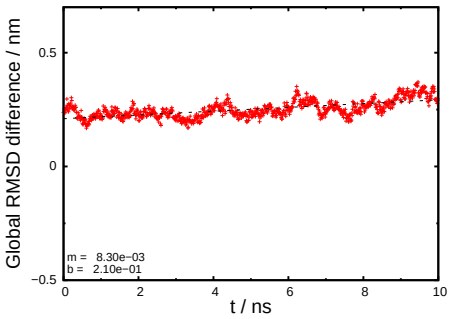
(c) 1.0 ns(b)



(d) 1.0 ns(c)

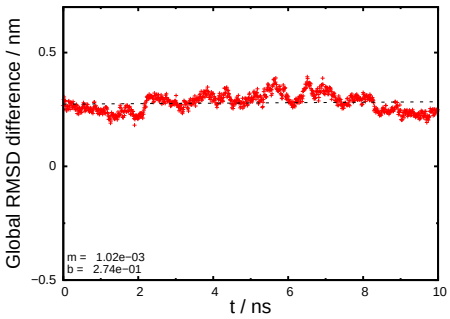


(e) 1.5 ns(a)

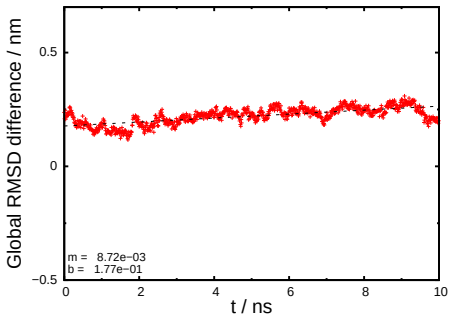


(f) 1.5 ns(b)

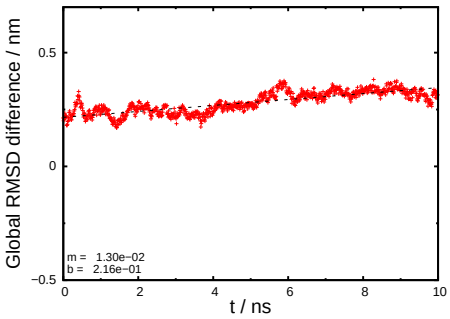
Figure D.2: See caption in **Figure D.5**.



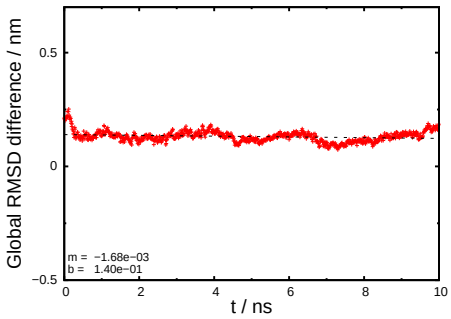
(a) 1.5 ns(c)



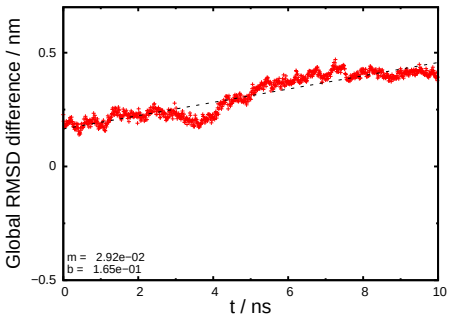
(b) 2.0 ns(a)



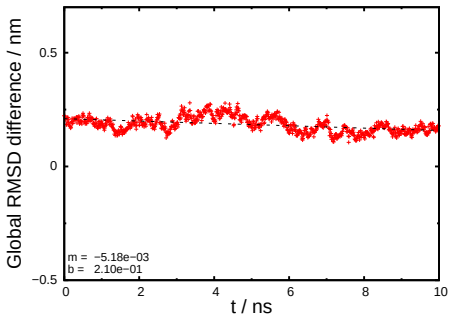
(c) 2.0 ns(b)



(d) 2.0 ns(c)

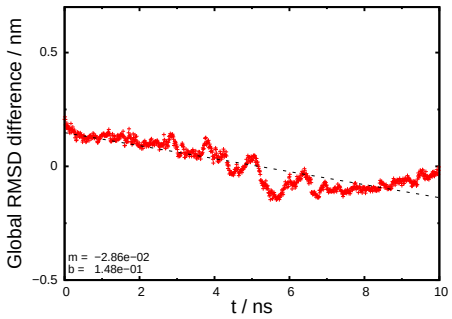


(e) 3.0 ns(a)

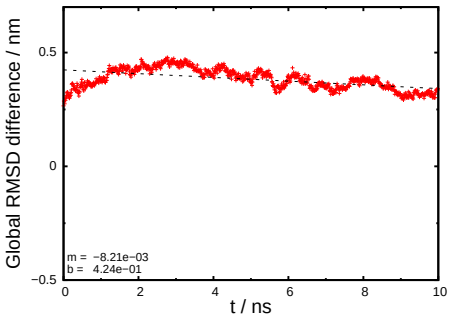


(f) 3.0 ns(b)

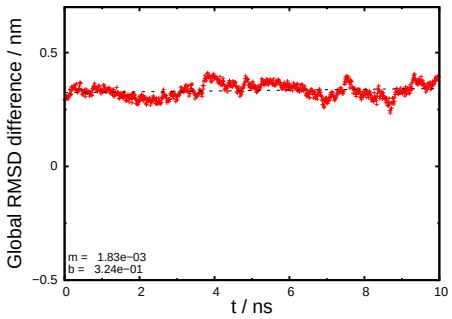
Figure D.3: See caption in **Figure D.5**.



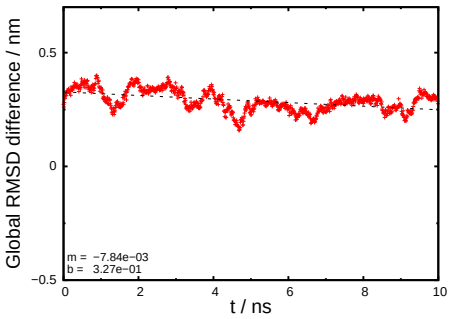
(a) 3.0 ns(c)



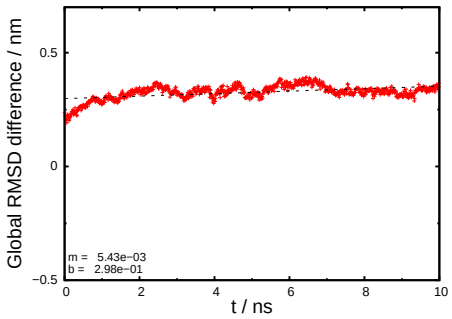
(b) 5.0 ns(a)



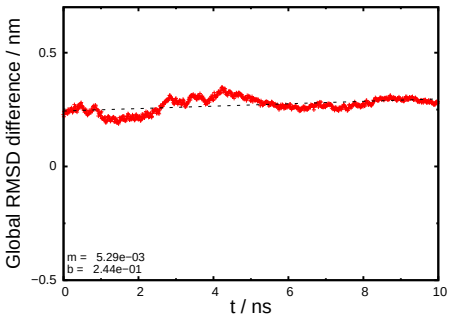
(c) 5.0 ns(b)



(d) 5.0 ns(c)

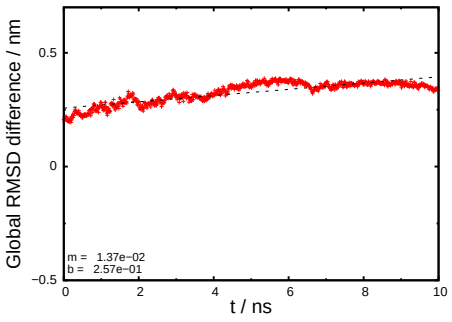


(e) 8.0 ns(a)



(f) 8.0 ns(b)

Figure D.4: See caption in **Figure D.5**.



(a) 8.0 ns(c)

Figure D.5: Difference of average RMSD value vs pH2mis and pH7mis conformations in the short simulations (**+**) and trendline (**- -**).

Appendix E

Energy landscapes and projections of short simulations

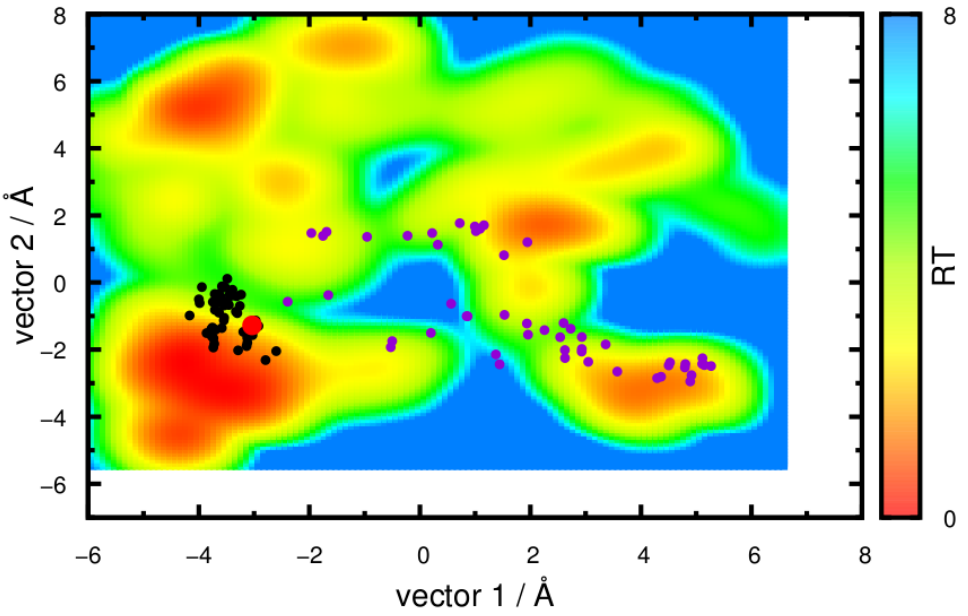


Figure E.1: Simulation: 0.5a. See caption in **Figure E.21**.

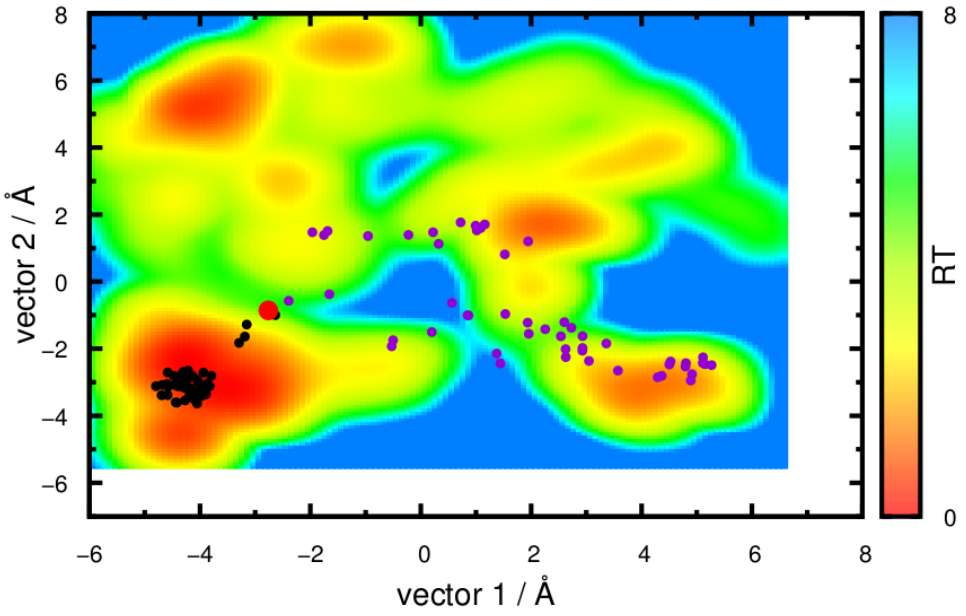


Figure E.2: Simulation: 0.5b. See caption in **Figure E.21**.

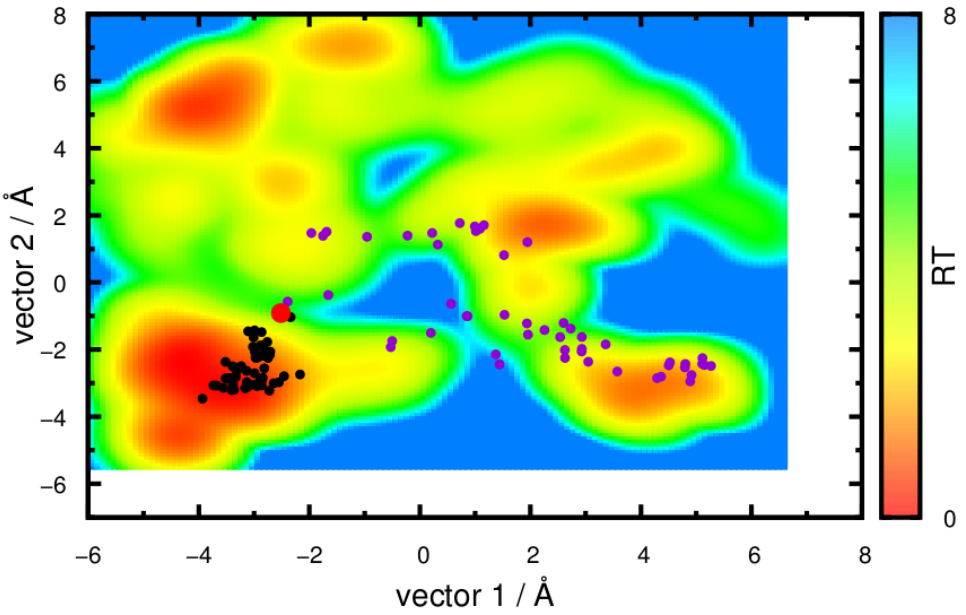


Figure E.3: Simulation: 0.5c. See caption in **Figure E.21**.

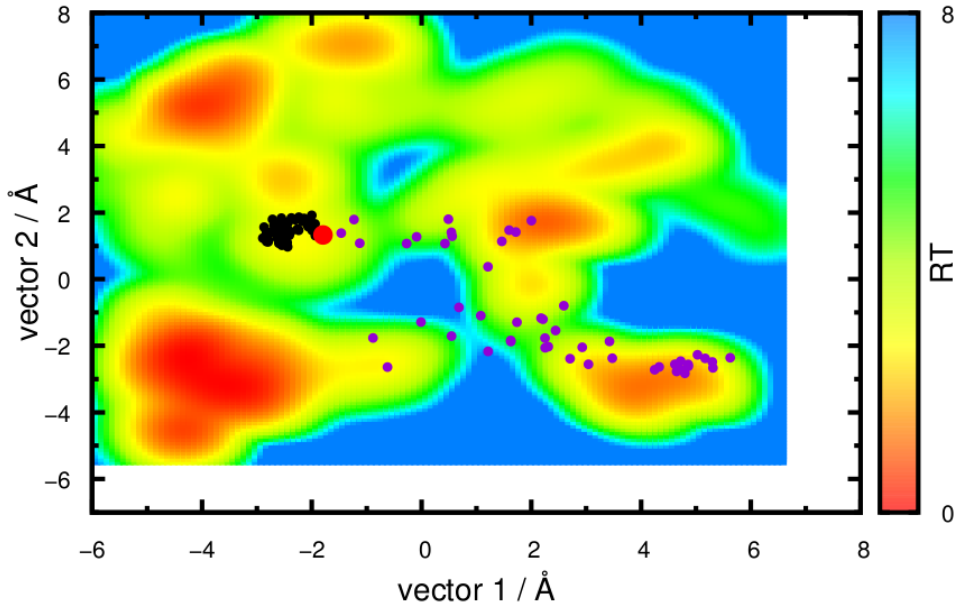


Figure E.4: Simulation: 1.0a. See caption in **Figure E.21**.

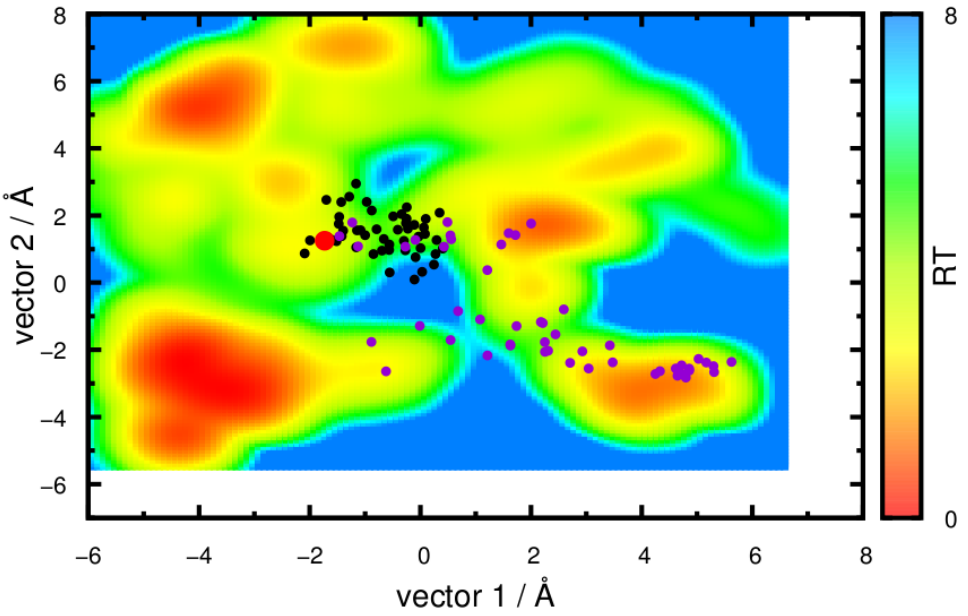


Figure E.5: Simulation: 1.0b. See caption in **Figure E.21**.

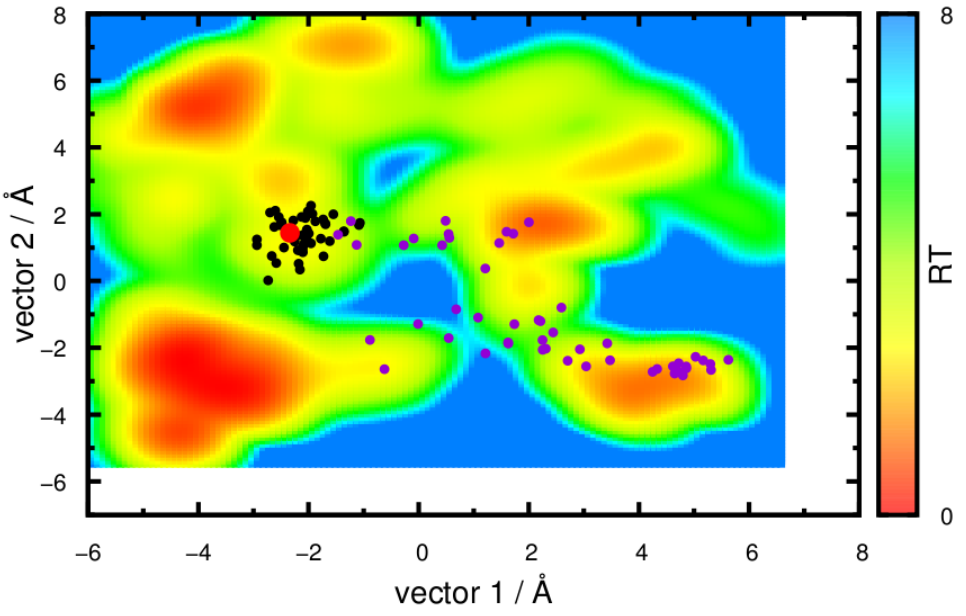


Figure E.6: Simulation: 1.0c. See caption in **Figure E.21**.

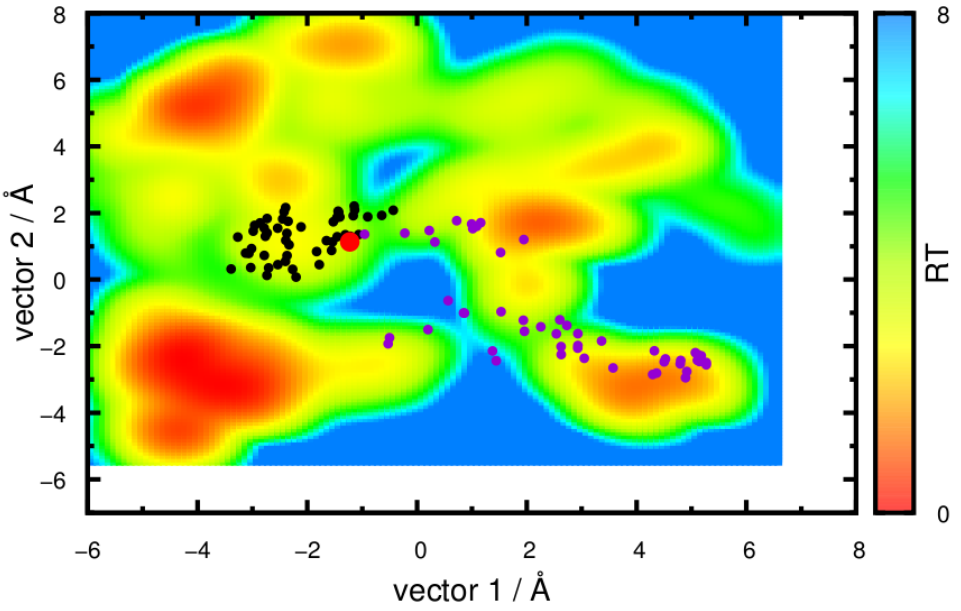


Figure E.7: Simulation: 1.5a. See caption in **Figure E.21**.

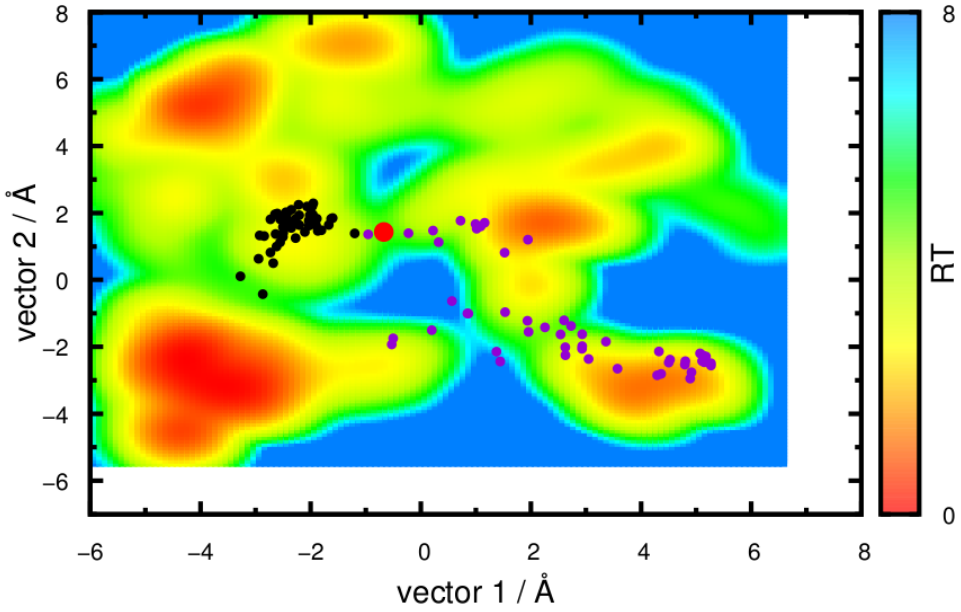


Figure E.8: Simulation: 1.5b. See caption in **Figure E.21**.

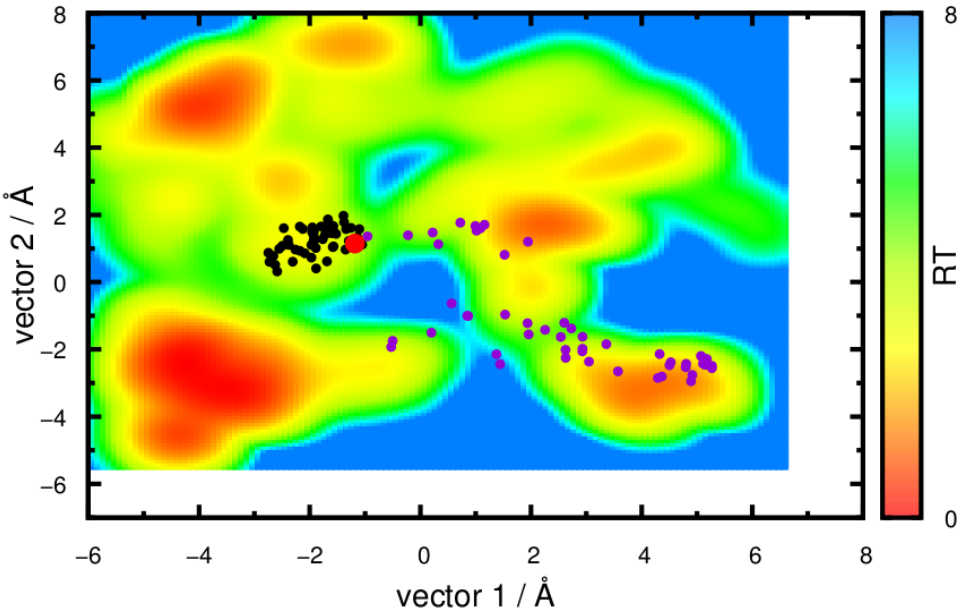


Figure E.9: Simulation: 1.5c. See caption in **Figure E.21**.

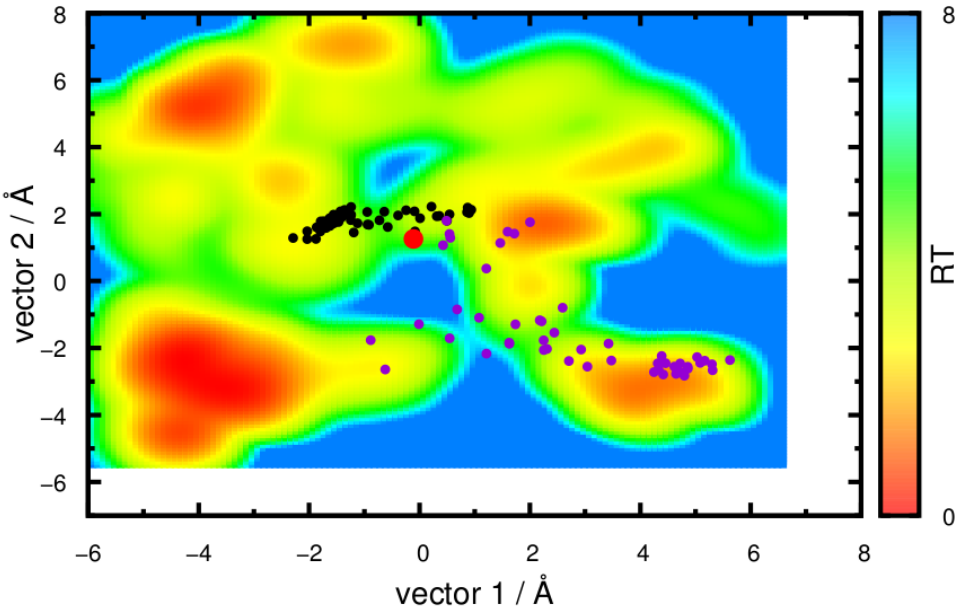


Figure E.10: Simulation: 2.0a. See caption in **Figure E.21**.

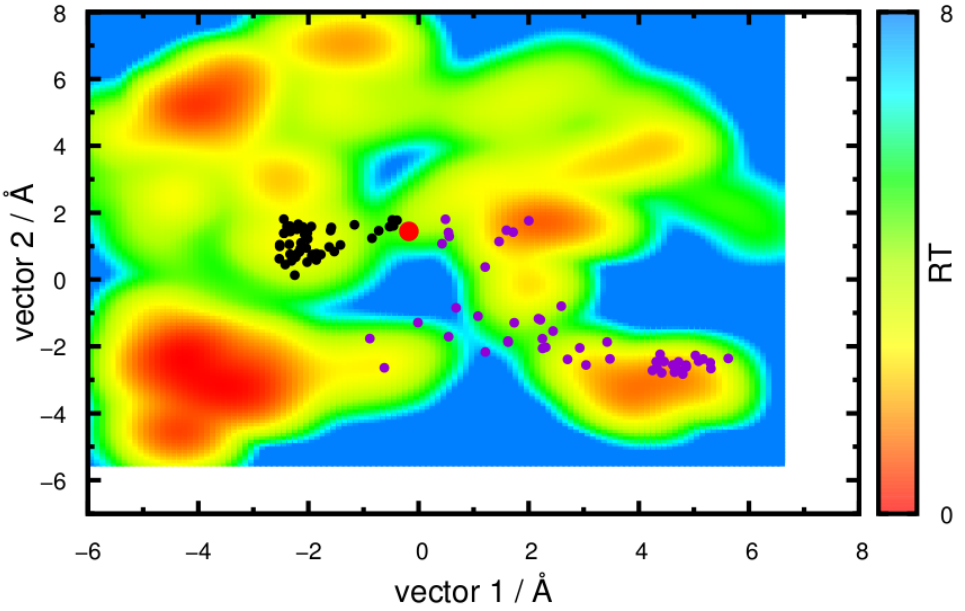


Figure E.11: Simulation: 2.0b. See caption in **Figure E.21**.

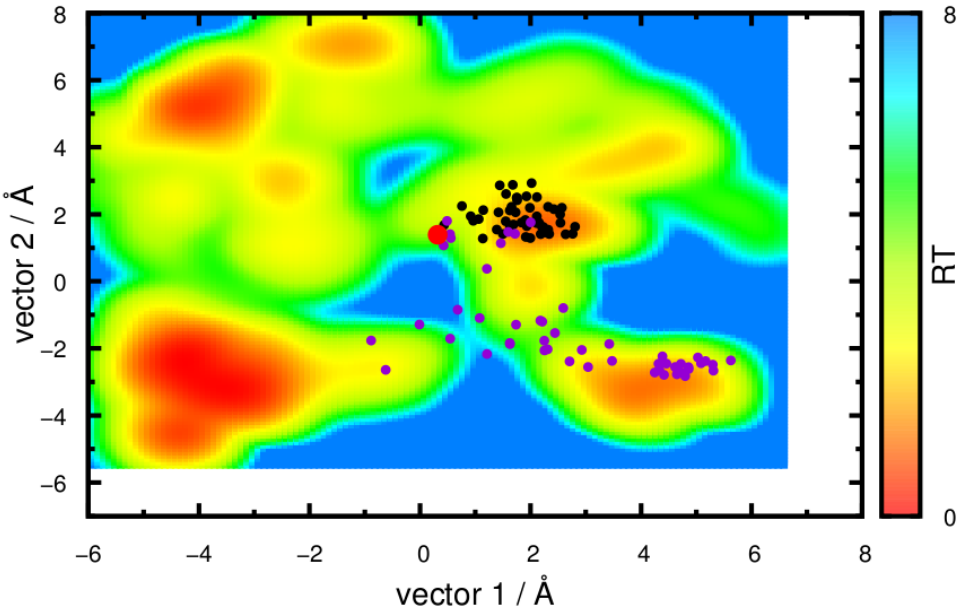


Figure E.12: Simulation: 2.0c. See caption in **Figure E.21**.

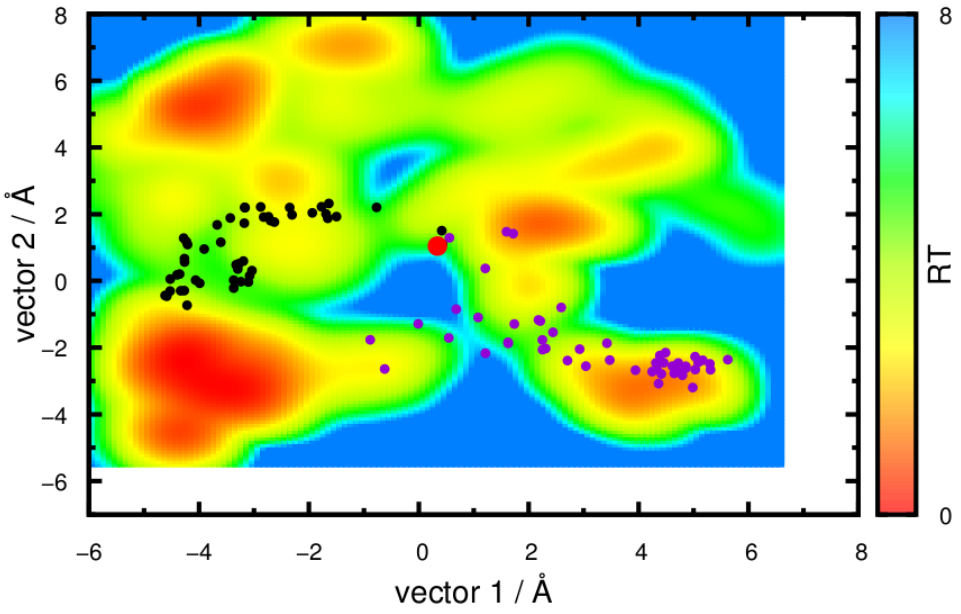


Figure E.13: Simulation: 3.0a. See caption in **Figure E.21**.

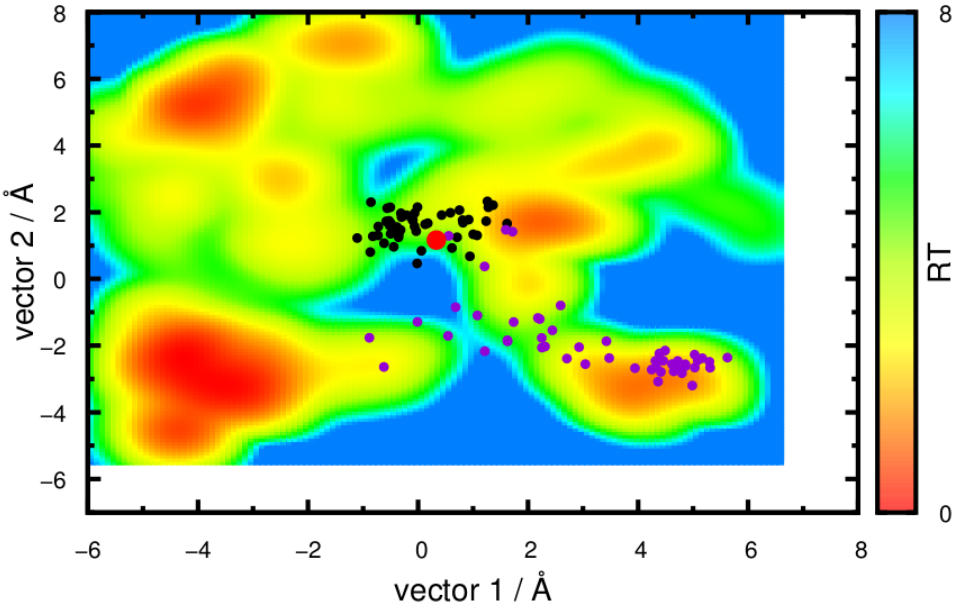


Figure E.14: Simulation: 3.0b. See caption in **Figure E.21**.

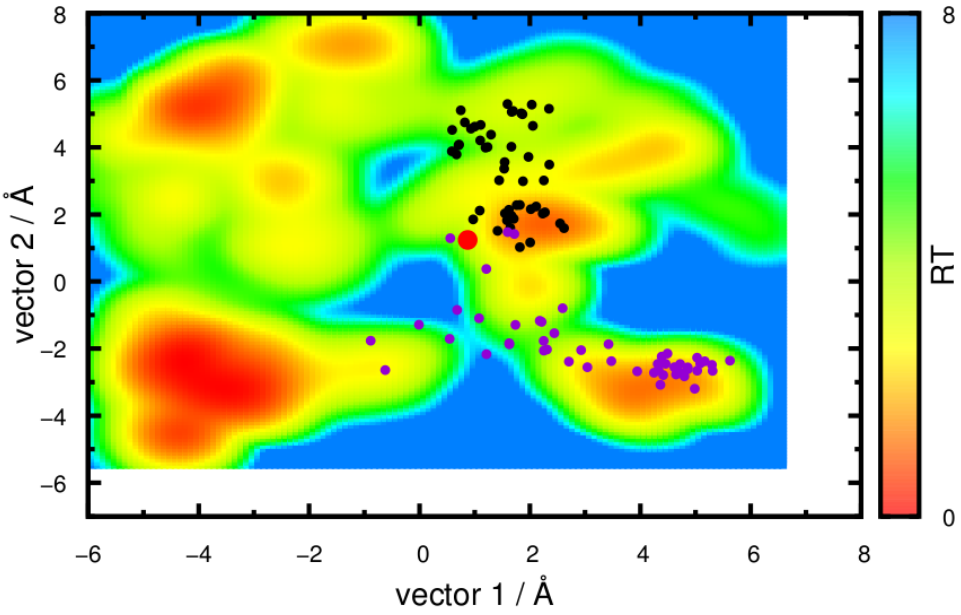


Figure E.15: Simulation: 3.0c. See caption in **Figure E.21**.

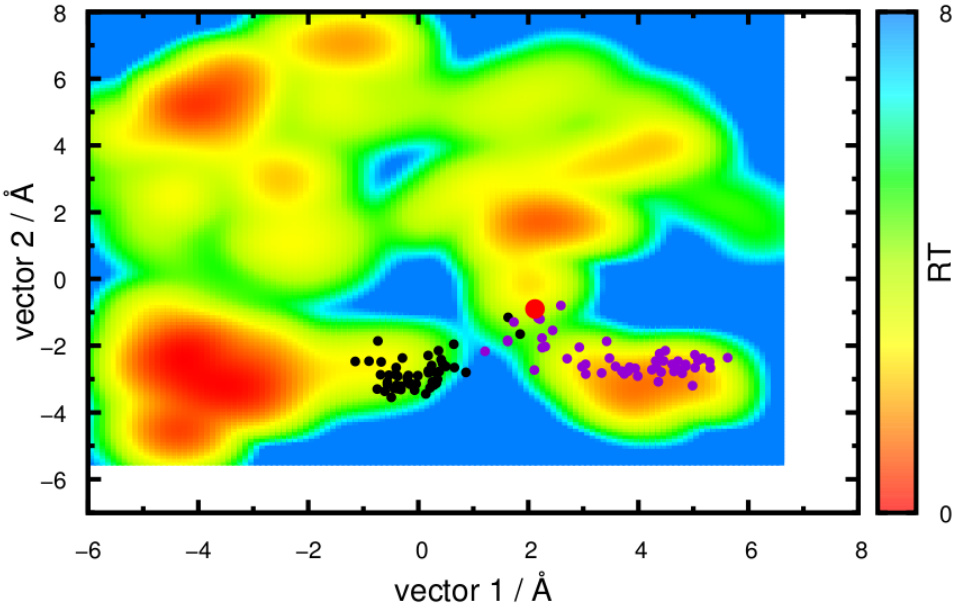


Figure E.16: Simulation: 5.0a. See caption in **Figure E.21**.

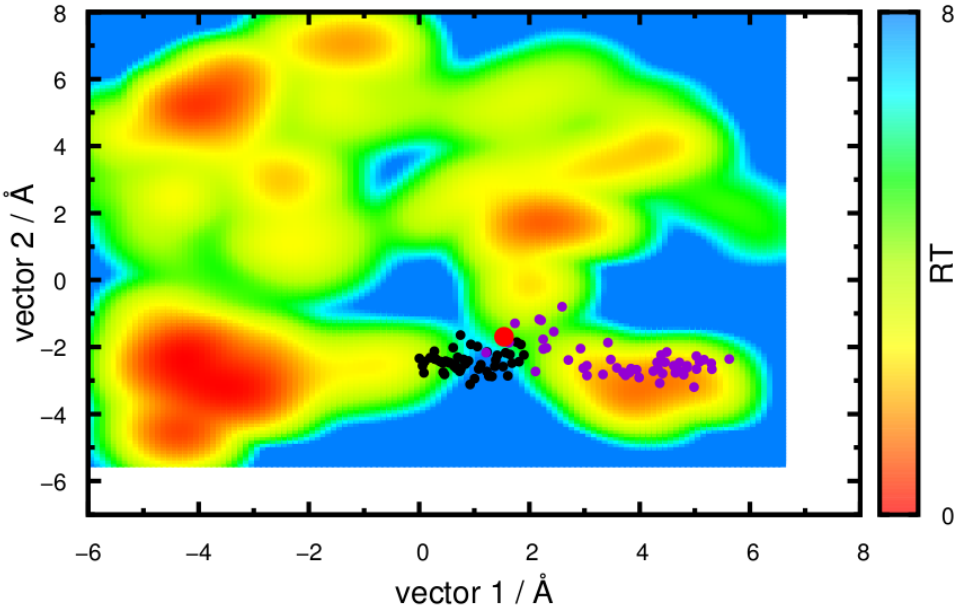


Figure E.17: Simulation: 5.0b. See caption in **Figure E.21**.

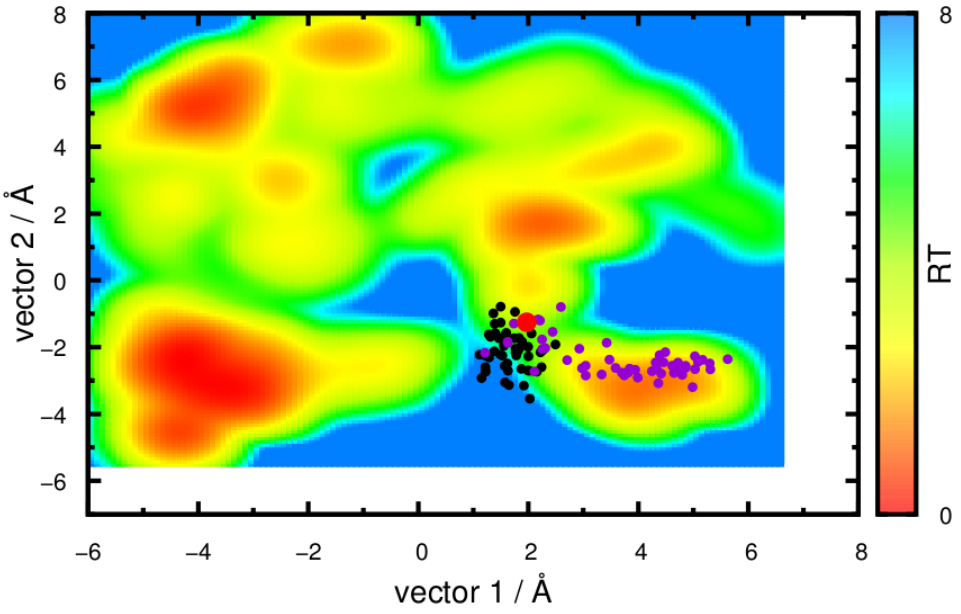


Figure E.18: Simulation: 5.0c. See caption in **Figure E.21**.

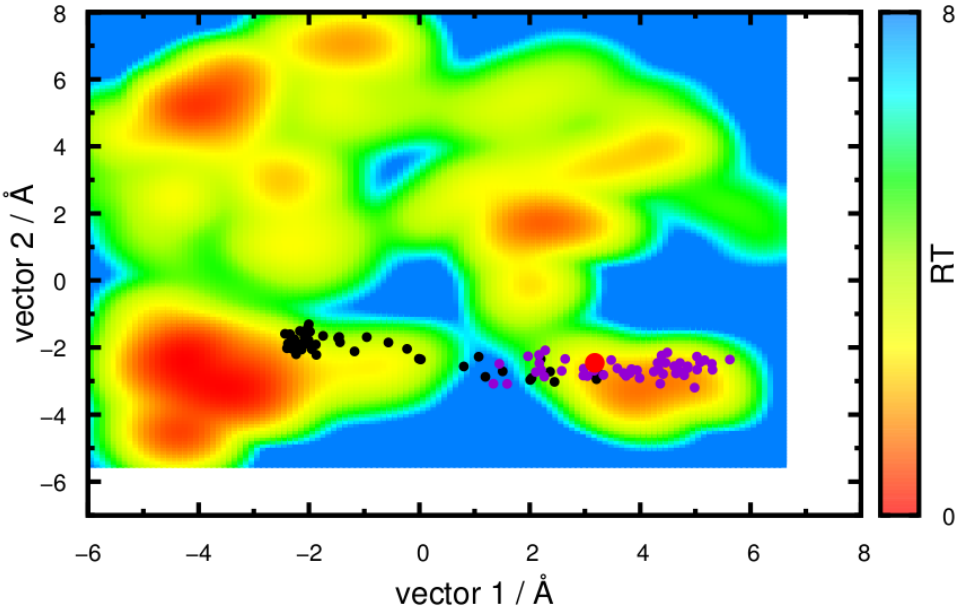


Figure E.19: Simulation: 8.0a. See caption in **Figure E.21**.

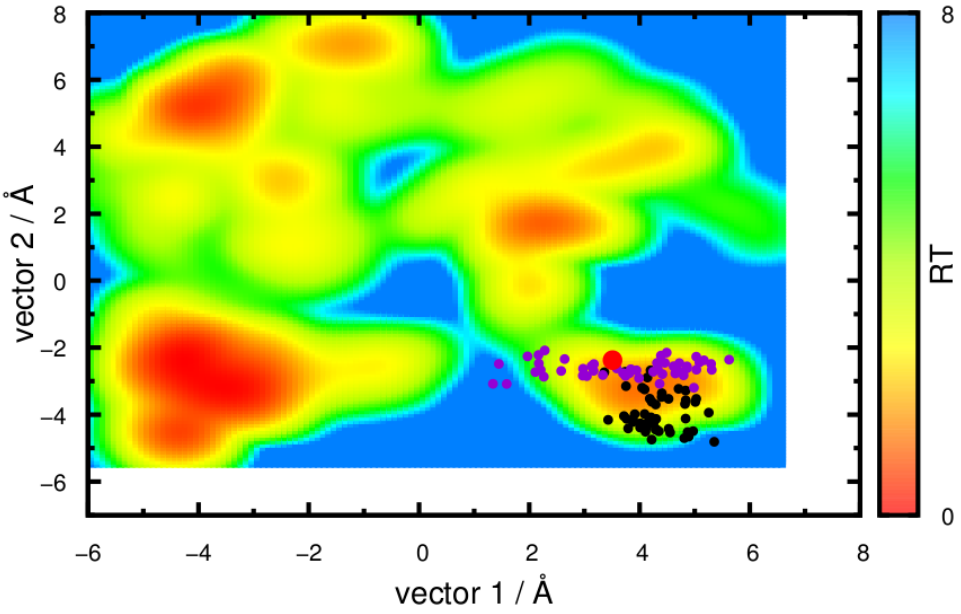


Figure E.20: Simulation: 8.0b. See caption in **Figure E.21**.

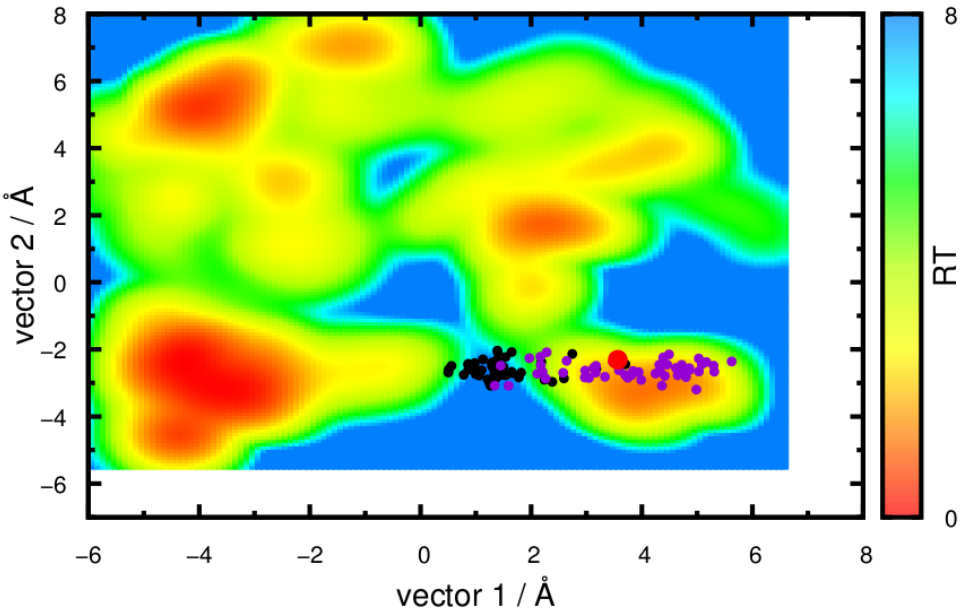


Figure E.21: Simulation: 8.0c. Energy landscapes at pH 7 using the first two PC's and projections of short simulations (●) and equivalent snapshots of the pH2mis simulation (●). Starting points of short simulations are also marked (●). Simulations are marked in captions.

Bibliography

[1] S R R Campos, M Machuqueiro, and A M Baptista. Constant-pH molecular dynamics simulations reveal a β-rich form of the human prion protein at low pH. *J Phys Chem B*, 2010. *in press*.

[2] M L DeMarco and V Daggett. Molecular mechanism for low pH triggered misfolding of the human prion protein. *Biochemistry*, 46:3045, 2007.

[3] G A Petsko and D Ringe. *Protein structure and function*. Blackwell Pub, 2004.

[4] D L Nelson and M M Cox. *Lehninger principles of biochemistry*. 2005.

[5] J C Kendrew, G Bodo, H M Dintzis, R G Parrish, H Wyckoff, and D C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662, 1958.

[6] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28:235, 2000.

[7] A G Murzin, S E Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536, 1995.

[8] J M Sanchez-Ruiz. Protein kinetic stability. *Biophys Chem*, 148:1, 2010.

[9] C K Mathews, K E van Holde, and K G Ahern. *Biochemistry*. Cummings, Redwood City, CA, 1990.

[10] D Whitford. *Proteins: structure and function*. Wiley, 2005.

[11] A M Lesk. *Introduction to protein science: architecture, function and genomics*. Oxford University Press, USA, 2004.

[12] G Forloni, L Terreni, I Bertani, S Fogliarino, R Invernizzi, A Assini, G Ribizzi, A Negro, E Calabrese, MA Volonté, et al. Protein misfolding in Alzheimer's and Parkinson's disease: genetics and molecular mechanisms. 23:957, 2002.

[13] F E Cohen and S B Prusiner. Pathologic conformations of prion proteins. *Annu Rev Biochem*, 67:793, 1998.

[14] S B Prusiner. Novel proteinaceous infectious particles cause scrapie. *Science*, 216:136, 1982.

[15] S Lehmann. Metal ions and prion diseases. *Curr Opin Chem Biol*, 6:187, 2002.

[16] S Mouillet-Richard, M Ermonval, C Chebassier, J L Laplanche, S Lehmann, J M Launay, and O Kellermann. Signal transduction through prion protein. *Science*, 289:1925, 2000.

[17] R Chiesa and D A Harris. Fishing for prion protein function. *PLoS Biol*, 7:439, 2009.

[18] W L DeLano. The PyMOL molecular graphics system, 2002.

[19] R Zahn, A Liu, T Lührs, R Riek, C von Schroetter, F L García, M Billeter, L Calzolai, G Wider, and K Wüthrich. NMR solution structure of the human prion protein. *Proc Natl Acad Sci USA*, 97:145, 2000.

[20] L Calzolai and R Zahn. Influence of pH on NMR structure and stability of the human prion protein globular domain. *J Biol Chem*, 278:35592, 2003.

[21] M L DeMarco and V Daggett. Local environmental effects on the structure of the prion protein. *C R Biologies*, 328:847, 2005.

[22] S B Prusiner. Prions. *Proc Natl Acad Sci USA*, 95:13363, 1998.

[23] J Collinge. Prion diseases of humans and animals: their causes and molecular basis. *Annu Rev Neurosci*, 24:519, 2001.

[24] A Aguzzi and M Polymenidou. Mammalian prion biology: one century of evolving concepts. *Cell*, 116:313, 2004.

[25] C Weissmann. Birth of a prion: spontaneous generation revisited. *Cell*, 122:165, 2005.

[26] K Abid and C Soto. The intriguing prion disorders. *Cell Mol Life Sci*, 63:2342, 2006.

[27] W Q Zou and P Gambetti. Prion: the chameleon protein. *Cell Mol Life Sci*, 64:3266, 2007.

[28] B Caughey, G S Baron, B Chesebro, and M Jeffrey. Getting a grip on prions: oligomers, amyloids, and pathological membrane interactions. *Annu Rev Biochem*, 78:177, 2009.

[29] N J Cobb and W K Surewicz. Prion diseases and their biochemical mechanisms. *Biochemistry*, 48:2574, 2009.

[30] K M Pan, M Baldwin, J Nguyen, M Gasset, A Serban, D Groth, I Mehlhorn, Z Huang, R J Fletterick, F E Cohen, and S B Prusiner. Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc Natl Acad Sci USA*, 90(23):10962, 1993.

[31] J E Arnold, C Tipler, L Laszlo, J Hope, M Landon, and R J Mayer. The abnormal isoform of the prion protein accumulates in late-endosome-like organelles in scrapie-infected mouse brain. *J Pathol*, 176:403, 1995.

[32] G S Jackson, L L P Hosszu, A Power, A F Hill, J Kenney, H Saibil, C J Craven, J P Waltho, A R Clarke, and J Collinge. Reversible conversion of monomeric human prion protein between native and fibrilogenic conformations. *Science*, 283:1935, 1999.

[33] R Gerber, A Tahiri-Alaoui, P J Hore, and W James. Conformational pH dependence of intermediate states during oligomerization of the human prion protein. *Protein Sci*, 17:537, 2008.

[34] D C Jenkins, D S Pearson, A Harvey, I D Sylvester, M A Geeves, and T J T Pinheiro. Rapid folding of the prion protein captured by pressure-jump. *Eur Biophys J*, 38:625, 2009.

[35] D O V Alonso, S J DeArmond, F E Cohen, and V Daggett. Mapping the early steps in the pH-induced conformational conversion of the prion protein. *Proc Natl Acad Sci USA*, 98:2985, 2001.

[36] E El-Bastawissy, M H Knaggs, and I H Gilbert. Molecular dynamics simulations of wild-type and point mutation human prion protein at normal and elevated temperature. *J Mol Graphics Modell*, 20:145, 2001.

[37] E Langella, R Improta, and V Barone. Checking the pH-induced conformational transition of prion protein by molecular dynamics simulations: effect of protonation of histidine residues. *Biophys J*, 87:3623, 2004.

[38] T L James, H Liu, N B Ulyanov, S Farr-Jones, H Zhang, D G Donnes, K Kaneko, D Groth, I Mehlhorn, S B Prusiner, and F E Cohen. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc Natl Acad Sci USA*, 94:10086, 1997.

[39] W E Morrell and J H Hildebrand. The distribution of molecules in a model liquid. *J Chem Phys*, 4:224, 1936.

[40] M P Allen and D J Tildesley. *Computer simulation of liquids*. Oxford University Press, USA, 1990.

[41] N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller, and E Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21:1087, 1953.

[42] A Rahman and F H Stillinger. Molecular dynamics study of liquid water. *J Chem Phys*, 55:3336, 1971.

[43] J A McCammon, B R Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267:585, 1977.

[44] E Lindahl and O Edholm. Mesoscopic undulations and thickness fluctuations in lipid bilayers from molecular dynamics simulations. *Biophys J*, 79:426, 2000.

[45] P L Freddolino, A S Arkhipov, S B Larson, A McPherson, and K Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14:437, 2006.

[46] D Bashford and M Karplus. pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*, 29:10219, 1990.

[47] M Feig, C L Brooks, et al. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr Opin Struct Biol*, 14:217, 2004.

[48] A Warshel. Calculations of chemical processes in solutions. *J Phys Chem*, 83:1640, 1979.

[49] A Warshel, F Sussman, and G King. Free energy of charges in solvated proteins: microscopic calculations using a reversible charging process. *Biochemistry*, 25:8368, 1986.

[50] K A Sharp and B Honig. Electrostatic interactions in macromolecules: theory and applications. *Annu Rev Biophys Biophys Chem*, 19:301, 1990.

[51] M K Gilson. Theory of electrostatic interactions in macromolecules. *Curr Opin Struct Biol*, 5:216, 1995.

[52] D Bashford. Macroscopic electrostatic models for protonation states in proteins. *Front Biosci*, 9:1082, 2004.

[53] A M Baptista. *Theoretical methods for the simulation of proteins at constant pH.* PhD thesis, Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Lisboa, 1998.

[54] A R Leach. *Molecular modelling: principles and applications.* Addison-Wesley Longman Ltd, 2001.

[55] O M Becker, A D MacKerell, B Roux, and M Watanabe, editors. *Computational biochemistry and biophysics.* Marcel Dekker, New York, 2001.

[56] A M Baptista, V H Teixeira, and C M Soares. Constant-pH molecular dynamics using stochastic titration. *J Chem Phys*, 117:4184, 2002.

[57] M Machuqueiro and A M Baptista. Constant-pH Molecular Dynamics with Ionic Strength Effects: Protonation-Conformation Coupling in Decalysine. *J Phys Chem B*, 110:2927, 2006.

[58] M Machuqueiro and A M Baptista. The pH-dependent conformational states of kyotorphin: a constant-pH molecular dynamics study. *Biophys J*, 92:1836, 2007.

[59] S R R Campos and A M Baptista. Conformational analysis in a multidimensional energy landscape: study of an arginylglutamate repeat. *J Phys Chem B*, 113:15989, 2009.

[60] P Magalhães. *Interaction between the neuropeptide kyotorphin and a lipid bilayer: a constant-pH Molecular Dynamics study.* Lisboa, 1998. Master Thesis.

[61] M Machuqueiro and A M Baptista. Acidic range titration of HEWL using a constant-pH molecular dynamics method. *Proteins: Struct Funct Bioinf*, 72:289, 2008.

[62] M Machuqueiro and A M Baptista. Molecular Dynamics at Constant pH and Reduction Potential: Application to Cytochrome c_3. *J Am Chem Soc*, 131:12586, 2009.

[63] J E Mertz and B M Pettitt. Molecular dynamics at a constant pH. *Int J High Perform Comput Appl*, 8:47, 1994.

[64] A M Baptista, P J Martel, and S B Petersen. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins: Struct Funct Bioinf*, 27:523, 1997.

[65] A M Walczak and J M Antosiewicz. Langevin dynamics of proteins at constant pH. *Phys Rev E*, 66:51911, 2002.

[66] M Dlugosz and J Antosiewicz. Constant-pH molecular dynamics simulations: a test case of succinic acid. *Chem Phys*, 302:161, 2004.

[67] J Mongan, D A Case, and J A McCammon. Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem*, 25:2038, 2004.

[68] U Börjesson and P H Hünenberger. Explicit-solvent molecular dynamics simulation at constant pH: methodology and application to small amines. *J Chem Phys*, 114:9706, 2001.

[69] A M Baptista. Comment on: explicit-solvent molecular dynamics simulation at constant pH: methodology and application to small amines by U Börjesson et al. *J Chem Phys*, 116:7766, 2002.

[70] M S Lee, F R Salsbury, and C L Brooks. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins*, 56:738, 2004.

[71] J Khandogin and CL Brooks. Constant pH molecular dynamics with proton tautomerism. *Biophys J*, 89:141, 2005.

[72] X Daura, A E Mark, and W F van Gunsteren. Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *J Comput Chem*, 19:535, 1998.

[73] W F van Gunsteren, S R Billeter, A A Eising, P H Hünenberger, P Krüger, A E Mark, W R P Scott, and I G Tironi. *Biomolecular Simulation: the GROMOS96 Manual and User Guide*. vdf Hochschulverlag AG an der ETH Zürich, Zürich, 1996.

[74] C Oostenbrink, A Villa, A E Mark, and W F van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53a5 and 53a6. *J Comput Chem*, 25:1656, 2004.

[75] W F van Gunsteren, X Daura, and A E Mark. *GROMOS force field*. Encyclopedia of Computational Chemistry, 1998.

[76] W R P Scott, P H Hünenberger, I G Tironi, A E Mark, S R Billeter, J F, A E Torda, T Huber, P Krüger, and W F van Gunsteren. The GROMOS biomolecular simulation program package. *J Phys Chem A*, 103:3596, 1999.

[77] C Oostenbrink, T A Soares, N F A van der Vegt, and W F van Gunsteren. Validation of the 53A6 GROMOS force field. *Eur Biophys J*, 34:273, 2005.

[78] A Villa, H Fan, T Wassenaar, and A E Mark. How sensitive are nanosecond molecular dynamics simulations of proteins to changes in the force field? *J Phys Chem B*, 111:6015, 2007.

[79] D Matthes and B L de Groot. Secondary structure propensities in peptide folding simulations: A systematic comparison of molecular mechanics interaction schemes. *Biophys J*, 97:599, 2009.

[80] R H Byrd, P Lu, J Nocedal, and C Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*, 16:1190, 1995.

[81] H J C Berendsen, J P M Postma, and W F van Gunsteren. Molecular dynamics and protein structure. *Polycrystal Book Series)(PO Box 27, Western Springs)*, page 18, 1985.

[82] I G Tironi, R Sperb, P E Smith, and W F van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *J Chem Phys*, 102:5451, 1995.

[83] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *J Chem Phys*, 81:3684, 1984.

[84] B Hess, H Bekker, H J C Berendsen, and J G E M Fraaije. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*, 18:1463, 1997.

[85] J P Ryckaert, G Ciccotti, and H J C Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys*, 23(3):327, 1977.

[86] S Miyamoto and P A Kollman. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem*, 13:952, 1992.

[87] F Fogolari, A Brigo, and H Molinari. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit*, 15:377, 2002.

[88] A M Baptista and C M Soares. Some theoretical and computational aspects of the inclusion of proton isomerism in the protonation equilibrium of proteins. *J Phys Chem B*, 105:293, 2001.

[89] J Hermans, H J C Berendsen, W F van Gunsteren, and J P M Postma. A consistent empirical potential for water-protein interactions. *Biopolymers*, 23:1513, 1984.

[90] D Bashford and K Gerwert. Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin. *J Mol Biol*, 224:473, 1992.

[91] H J C Berendsen, D van der Spoel, and R van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput Phys Commun*, 91:43, 1995.

[92] E Lindahl, B Hess, and D van der Spoel. GROMACS 30: a package for molecular simulation and trajectory analysis. *J Mol Model*, 7:306, 2001.

[93] D van der Spoel, E Lindahl, B Hess, G Groenhof, A E Mark, and H J C Berendsen. GROMACS: fast, flexible, and free. *J Comput Chem*, 26:1701, 2005.

[94] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577, 1983.

[95] I T Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, second edition, 2002.

[96] B W Silverman. *Density estimation for statistcs and data analysis*. Chapman and Hall, London, 1986.